CONDITIONAL DEPENDENCE VIA SHANNON CAPACITY: AXIOMS, ESTIMATORS AND APPLICATIONS

BY

WEIHAO GAO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Advisers:

Assistant Professor Sewoong Oh
Professor Pramod Viswanath

ABSTRACT

We consider axiomatically the problem of estimating the strength of a conditional dependence relationship $P_{Y|X}$ from a random variable X to a random variable Y. This has applications in determining the strength of a known causal relationship, where the strength depends only on the conditional distribution of the effect given the cause (and not on the driving distribution of the cause). Shannon capacity, appropriately regularized, emerges as a natural measure under these axioms. We examine the problem of calculating Shannon capacity from the observed samples and propose a novel fixed-k nearest-neighbor estimator, and demonstrate its consistency. Finally, we demonstrate an application to single-cell flow-cytometry, where the proposed estimators significantly reduce sample complexity.

To my parents, for their love and support.

ACKNOWLEDGMENTS

I would first like to thank my thesis co-advisors Prof. Sewoong Oh and Prof. Pramod Viswanath, without whom this thesis would not have been possible. Their sincere and patient guidance and support led me to the world of research, that benefited both my research projects and career endeavors. I also want to thank Prof. Sreeram Kannan, of the University of Washington, who contributed advice on this thesis.

I would like to thank Prof. Yuliy Baryshnikov, Prof. Maxim Raginsky and Prof. Yihong Wu (now at Yale University) for serving as my qualifying exam committee. I would not be able to pursue my Ph.D. without their help. I would like to thank all the faculty who taught me during the past two years, including Prof. Chandra Chekuri, Prof. Xiaochun Li, Prof. Pierre Moulin, Prof. Sewoong Oh, Prof. Maxim Raginsky, Prof. Zhongjin Ruan, Prof. Renming Song, Prof. Rayadurgam Srikant and Prof. Venugopal V. Veeravalli. I also want to thank all the TAs in my courses for their work.

Finally, I would like to thank my parents and family for providing me with support and encouragement throughout my years of study. Without their support from across the Pacific Ocean, this thesis would not have been possible. To them I dedicate this thesis.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The axiomatic study of dependence measures on joint distributions between two random variables X and Y has a long history in statistics [1, 2, 3]. In this thesis, we study the relatively unexplored terrain of measures that depend only on the conditional distribution $P_{Y|X}$. We are motivated to study conditional dependence measures by a problem in causal strength estimation. Causal learning is a basic problem in many areas of scientific learning, where one wants to uncover the cause-effect relationship using interventions or directly from observational data [4, 5, 6]. In this thesis, we are interested in an even simpler question: Given a causal relationship, how does one measure the strength of the relationship? This problem arises in many contexts, for example, one may know causal genetic pathways but only a subset of these may be active in a particular tissue or organ — therefore, deducing how much influence each causal link exerts becomes necessary.

We focus on a simple model: consider a pair of random variables (X, Y) with *known* causal direction $X \to Y$, and suppose that there are no confounders. We denote the causal influence quantity by $\mathcal{C}(X \to Y)$. There are two philosophically distinct ways to model the quantity: the first one is *factual influence*, i.e., how much influence X exerts on Y under the current probability of the cause X. The second possible way, which one can term as *potential influence* measures models how much influence X can exert on Y in principle — without cognizance to the present distribution of the cause. For example, consider a city which has very few smokers, but smoking inevitably leads to lung cancer. In such a city, the factual influence of smoking on lung cancer will be small but the potential influence is very high. Depending on the setting, one may prefer the former or the latter. In this thesis, we are interested in the *potential influence* of a cause on its effect.

We want $\mathcal{C}(X \to Y)$ to be invariant to scaling and one-one transformations of the variables X, Y. This naturally suggests information theoretic metrics as plausible choices of $\mathcal{C}(X \to Y)$, starting with the mutual information $I(X;Y) = D(P_{XY}||P_X P_Y)$, at least in the case of factual influence. This measures the information through the channel from $X \to Y$ as given by the prior P_X. Observe that this metric is *symmetric* with respect to the directions $X \to Y$ and $Y \to X$; this property is not always desirable. In fact, this measure is taken as a starting point to develop an axiomatic approach to studying causal strength on general graphs in [7].

Potential causal influence is posited as a relevant metric to spot "trends" in gene pathways in a recent work [8]. In the particular application considered, rare biological states of gene X in a given data source may nevertheless correspond to important biological states, and therefore it is important to have causal measures that are not sensitive to the cause distribution but only depend on the relationship between the cause and the effect. To quantify the potential influence of those rare values of X, the following approach is proposed. Replace the observed distribution P_X by a *uniform* distribution U_X and calculate the mutual information under the joint distribution $U_X P_{Y|X}$. The resulting causal strength quantification is $\mathcal{C}(X \to Y) = D(U_X P_{Y|X}||P_U P_Y)$, where P_Y represents the distribution at the output of a channel $P_{Y|X}$ with input given by U_X. We call this quantification Uniform Mutual Information (UMI). A key challenge is to compute this quantity from i.i.d. samples in a statistical efficient manner, especially when the channel output is continuous valued and potentially in high dimensions. This is the first focus point of the thesis.

UMI is not invariant under bijective transformations and is also sensitive to the estimated support size of X. Even more fundamentally, it is unclear why one would prefer the uniform prior distribution to measuring the potential influence through the channel $P_{Y|X}$. Based on natural axioms of data processing and additivity, we propose an alternative measure of causal strength: the *largest amount of information* that can be sent through the channel, namely the *Shannon capacity*. Formally $\mathcal{C}(X \to Y) = \max_{Q_X} D(Q_X P_{Y|X}||Q_X P_Y)$, where P_Y represents the distribution at the output of a channel $P_{Y|X}$ with input given by Q_X. We refer to such a quantification as Capacitated Mutual Information (CMI). A key challenge is to compute this quantity from i.i.d. samples in a statistically efficient manner, especially when the channel output is continuous valued and potentially in high dimensions. This is the second focus point of the thesis.

We make the following **main contributions** in this thesis.

- *UMI Estimation*: We construct a novel estimator to compute UMI from data sampled i.i.d. from a distribution P_{XY}. The estimator brings together ideas from three disparate threads in statistical estimation theory: nearest-neighbor methods, a correlation-boosting idea in the estimation of mutual information from samples [9], and importance sampling. The estimator has only a *single* hyper parameter (the number of nearest-neighbors considered), uses an off-the-shelf kernel density estimator of only P_X, and has strong connections to the entropy estimator of [10]. Our main technical result is to show that the estimator is consistent supposing that the Radon-Nikodym derivative $\frac{dP_U}{dP_X}$ is uniformly bounded over the support. In simulations, the estimator has shown very strong performance in terms of sample complexity compared to a baseline of the partition-based estimator in [11].

- *CMI Estimation*: We build upon the estimator derived for UMI and construct an optimization problem that mimics the optimization problem inherent in computing the capacity directly from the conditional probability distribution of the channel. Our main technical result is to show the consistency of this estimator, supposing that the Radon-Nikodym derivative $\frac{dP_Q}{dP_X}$ is uniformly bounded over the support, where P_Q is the optimizing input to the channel. Simulation results show strong empirical performance compared to a baseline of a partition-based method followed by discrete optimization.

- *Application to Gene Pathway Influence*: In [8], considered an important work in single-cell flow-cytometry data analysis, a causal strength metric (termed DREMI) is proposed for measuring the causal influence of a gene. This estimator is a specific way of implementing UMI along with a "channel amplification" step, and DREMI was successfully used to spot gene-pathway trends. We show that our proposed CMI and UMI estimators also exhibit the same performance as DREMI when supplied with the full dataset, while at the same time, they have significantly smaller sample complexity for the same performance.

CHAPTER 2

AN AXIOMATIC APPROACH

In this chapter, we are interested in a basic question: What properties should an influence measure satisfy? We will answer this question by formally modeling an influence measure on conditional probability distributions, by postulating five natural axioms.

2.1 Axioms on Influence Measures

Let X be drawn from an alphabet $\mathcal{X}$, and Y from an alphabet $\mathcal{Y}$. Let the probability distribution of Y given X be given as $P_{Y|X}$. Let $\mathcal{P}$ be a family of conditional distributions; usually we will consider the case when $\mathcal{P}$ is the set of all possible conditional distributions. Then the influence measure $\mathcal{C}(X \to Y)$ is a function of the conditional distribution to non-negative real numbers: $\mathcal{C} : \mathcal{P}(\mathcal{Y}|\mathcal{X}) \to \mathbb{R}^+$, and we can write $\mathcal{C}(X \to Y)$ as $\mathcal{C}(P_{Y|X})$. We postulate that the function $\mathcal{C}$ satisfies five axioms on $\mathcal{P}$, and show that CMI satisfies all five axioms:

0. **Independence**: The measure $\mathcal{C}(P_{Y|X}) = 0$ if and only if $P_{Y=y|X=x}$ depends only on y.

1. **Data Processing**: Let $X \to Y \to Z$ be a processing chain, i.e., $P_{Z=z|X=x} = \sum_{y \in \mathcal{Y}} P_{Z=z|Y=y} P_{Y=y|X=x}$, then the natural data processing inequalities should hold: (a) $\mathcal{C}(P_{Y|X}) \geq \mathcal{C}(P_{Z|X})$; and (b) $\mathcal{C}(P_{Z|Y}) \geq \mathcal{C}(P_{Z|X})$.

2. **Additivity**: For a parallel channel $P_{Y_1,Y_2|X_1,X_2} := P_{Y_1|X_1} P_{Y_2|X_2}$, we need

$$\mathcal{C}(P_{Y_1,Y_2|X_1,X_2}) = \mathcal{C}(P_{Y_1|X_1}) + \mathcal{C}(P_{Y_2|X_2}). \tag{2.1}$$

3. **Monotonicity**: A causal relationship is strong if many possible values of P_Y are achievable by varying the input probability distribution P_X. Thus if we consider $P_{Y|X}$ as a map from the probability simplex in X to the probability simplex in Y, the larger the range of this map, the stronger should be the causal strength.

    (a) $\mathcal{C}$ should only depend on the range of the map, $\text{Range}(P_{Y|X})$, the convex hull of the output distributions $P_{Y|X=x}$.

    (b) $\mathcal{C}$ should be a monotonic function of the range of the map. If $P_{Y|X}$ and $Q_{Y|X}$ are such that, $\text{Range}(P_{Y|X}) \subseteq \text{Range}(Q_{Y|X})$ then: $\mathcal{C}(P_{Y|X}) \leq \mathcal{C}(Q_{Y|X})$.

4. **Maximum Value**: The maximum value over all possible conditional distributions for a particular output alphabet $\mathcal{Y}$ should be achieved exactly when the relationship is fully causal, i.e., each $Y = y$ can be achieved by setting $X = x$ for some x.

We begin our exploration of appropriate influence measures with the alphabets for X and Y being discrete. Let $I(P_{XY}) := D(P_{XY}||P_X P_Y)$ denote the mutual information with respect to the joint distribution P_{XY}. Since we are looking at *potential* influence measures, Shannon capacity, defined as the maximum over input probability distributions of the mutual information, is a natural choice:

$$\text{CMI}(P_{Y|X}) := \max_{P_X} I(P_X P_{Y|X}). \tag{2.2}$$

Our first claim is the following:

Proposition 1. *CMI satisfies all the axioms of causal influence.*

Proof: The proof is fairly straightforward.

- Clearly Axiom 0 holds, cf. Chapter 2 of [12].

- Axiom 1: Suppose $\text{CMI}(P_{Z|X})$ is achieved with P_X^*. Consider the joint distribution $P_X^* P_{Y|X} P_{Z|Y}$. Utilizing the data-processing inequality for mutual information, we get

$$\begin{aligned}\text{CMI}(P_{Y|X}) &= \max_{P_X} I(P_X P_{Y|X}) \geq I(P_X^* P_{Y|X}) \\ &\geq I(P_X^* P_{Z|X}) = \text{CMI}(P_{Z|X}).\end{aligned} \tag{2.3}$$

Thus Axiom 1a is satisfied. Now consider Axiom 1b. With the same joint distribution, let P_Y^* be the marginal of Y. Then we have,

$$\begin{aligned}\mathrm{CMI}(P_{Z|Y}) &= \max_{P_Y} I(P_Y P_{Z|Y}) \geq I(P_Y^* P_{Z|Y}) \\ &\geq I(P_X^* P_{Z|X}) = \mathrm{CMI}(P_{Z|X}). \end{aligned} \tag{2.4}$$

- Axiom 2: This is a standard result for Shannon capacity and we refer the interested reader to Chapter 7 of [12].

- Axiom 3a: First we rewrite capacity equivalently as the information-centroid (see [13]):

$$\begin{aligned}\mathrm{CMI}(P_{Y|X}) &:= \max_{P_X} \min_{q_Y} D(P_{Y|X} \| q_Y | P_X) \\ &= \min_{q_Y} \max_{P_X} D(P_{Y|X} \| q_Y | P_X) \\ &= \min_{q_Y} \max_{x} D(P_{Y|X=x} \| q_Y). \end{aligned} \tag{2.5}$$

  Here the conditional KL divergence $D(X \| Y | Z)$ is defined in the usual way:

$$D(X \| Y | Z) = \sum_z P_Z(z) \sum_{(x,y)} P_{X|Z}(x|z) \log \frac{P_{X|Z}(x|z)}{P_{Y|Z}(y|z)}. \tag{2.6}$$

  This characterization allows us to make the observation that the capacity is a function only of the convex hull of the probability distributions $P_{Y|X=x}$. Given a conditional probability distribution $P_{Y|X}$, we augment the input alphabet to have one more input symbol x' such that $P_{Y|X=x'} = \sum_x \alpha_x P_{Y|X=x}$ is a convex combination of the other conditional distributions. We claim that the capacity of the new channel is unchanged: one direction is obvious, i.e., the new channel has capacity greater than or equal to the original channel, since adding a new symbol cannot decrease capacity. To show the other direction, we use (2.5) and observe that, due to the convexity of KL divergence in its arguments, we get,

$$D(P_{Y|X=x'}\|q_Y) = D(\sum_x \alpha_x P_{Y|X=x}\|q_Y)$$
$$\leq \sum_x \alpha_x D(P_{Y|X=x}\|q_Y) \leq \max_x D(P_{Y|X=x}\|q_Y). \tag{2.7}$$

Thus Shannon capacity is only a function of the convex hull of the range of the map $P_{Y|X}$, satisfying Axiom $3a$. This function is monotonic directly from (2.5), thus satisfying Axiom $3b$.

- Axiom 4: For fixed output alphabet $\mathcal{Y}$, it is clear that $\max_{\mathcal{X},P_{Y|X}}$ CMI $=$ $\log|\mathcal{Y}|$. Now suppose for some conditional distribution $P_{Y|X}$ we have $\mathrm{CMI}(P_{Y|X}) = \log|\mathcal{Y}|$. This implies that, with the optimizing input distribution, $H(Y) - H(Y|X) = \log|\mathcal{Y}|$. This implies that $H(Y) = \log|\mathcal{Y}|$ and $H(Y|X) = 0$, thus Y is a deterministic function of the essential support of X and since $H(Y) = \log|\mathcal{Y}|$, it implies that $P_Y = U_Y$, the uniform distribution and the deterministic function is onto.

Axiomatic View of UMI: Now consider an alternative metric: Uniform Mutual Information (UMI) which is defined as the mutual information with uniform input distribution,

$$\mathrm{UMI}(P_{Y|X}) := I(U_X P_{Y|X}), \tag{2.8}$$

where U_X is the uniform distribution on $\mathcal{X}$. This estimator is motivated by the recent work in [8]. We investigate how this estimator fares in terms of the proposed axioms.

- UMI clearly satisfies Axiom 0. It also satisfies Axiom $1a$. Data-processing inequality for mutual information on the joint distribution $U_X P_{Y|X} P_{Z|Y}$ implies that $I(U_X P_{Y|X}) \geq I(U_X P_{Z|X})$, which is the same as $\mathrm{UMI}(P_{Y|X}) \geq \mathrm{UMI}(P_{Z|X})$. Thus $I(U_Y P_{Z|Y}) \geq I(U_X P_{Z|X})$.

- UMI however does not satisfy Axiom $1b$ in general. However, if the transition matrices $P_{Y|X}$ and $P_{Z|Y}$ are both doubly stochastic, then a straightforward calculation shows that UMI satisfies Axiom $1b$ too.

- UMI satisfies Axiom 2 since the uniform distribution on X_1, X_2 naturally factors as $U_{X_1,X_2} = U_{X_1}U_{X_2}$ and we have $\mathrm{UMI}(P_{Y_1,Y_2|X_1,X_2})$

$$\begin{aligned} &= I(U_{X_1,X_2}P_{Y_1,Y_2|X_1,X_2}) \\ &= I(U_{X_1}U_{X_2}P_{Y_1|X_1}P_{Y_2|X_2}) \\ &= \mathrm{UMI}(P_{Y_1|X_1}) + \mathrm{UMI}(P_{Y_2|X_2}). \end{aligned} \tag{2.9}$$

- UMI does not satisfy Axiom 3a since multiple repeated values of $P_{Y|X=x}$ does not alter the convex hull but alters the UMI value.
- Interestingly, UMI does satisfy Axiom 4 for the same reason as CMI.

2.2 Real-Valued Alphabets

For real-valued X, the Shannon mutual information is not finite without additional regularizations. This is also true of other measures of relation such as the Renyi correlation [2], and in each case the measure is studied in the context of some form penalty term. Typically this is done via a cost constraint on the real-valued input parameters. In this context, one possibility is to consider the following norm-constrained optimization to ensure the causal effect is finite valued:

$$\mathrm{CMI}(P_{Y|X}, a) := \max_{P_X : \mathbb{E}\|X\|_2^2 \le a} I(P_X P_{Y|X}). \tag{2.10}$$

In practice, a is chosen from the empirical moments of X from samples: $a := \frac{1}{N}\sum_{i=1}^{N} \|X_i\|_2^2$ for samples $X_1, \ldots, X_N$. This regularization turns out to be the so-called "power constraint" on the input, common in treatments of additive noise communication channels.

CHAPTER 3

ESTIMATORS

Although the definitions of UMI and CMI seamlessly apply to both discrete and continuous random variables, the estimation becomes relatively straightforward when both $\mathcal{X}$ and $\mathcal{Y}$ are discrete; the estimation of the conditional distribution $P_{Y|X}$ and the computation of UMI and CMI can be separated in a straightforward manner. For this reason and also due to an application in genomic biology that we study, we focus on the more challenging regime that $\mathcal{Y}$ is continuous. Due to certain subtleties in the estimation process, we provide separate estimators each customized for each case of discrete and continuous $\mathcal{X}$, respectively.

3.1 Uniform Mutual Information

The idea of applying UMI to infer the strength of conditional dependence was first proposed in [8]. Off-the-shelf two-dimensional kernel density estimators (KDE) are used to first estimate the joint distribution P_{XY} from given samples. Subsequently, the channel $P_{Y|X}$ is computed directly from the joint distribution, and then UMI is computed via either numerical integration or sampling from U_X and $P_{Y|X}$. This approach suffers from known drawbacks of KDE, such as sensitivity to the choice of the bandwidth and increased bias in higher-dimensional X and Y. However, a more critical challenge in using KDE to estimate the joint distribution at all points (and not just at samples) is the *overkill* nature: we only need to compute a single functional (UMI) of the joint distribution, which could in principle be computed more efficiently directly from the samples. It is not at all clear how to *directly* estimate UMI.

Perhaps surprisingly, we bring together ideas from three topics in statistical estimation to introduce novel estimators that are also provably convergent. Our estimator is based on (a) k-nearest-neighbor estimators, e.g. [10]; (b) the correlation boosting idea of the estimator from [9], which is widely adopted in practice [14]; and (c) the importance-sampling techniques to adjust for the uniform prior for UMI. We explain each idea below.

Consider a simpler task of computing the mutual information from samples; several approaches exist for this estimation: [15, 9, 16, 17, 18, 19, 20, 21, 22]. Note that three applications of the entropy estimator, such as those from [23], give an estimate of the mutual information, i.e. $\widehat{I}(X;Y) = \widehat{H}(X) + \widehat{H}(Y) - \widehat{H}(X,Y)$. Each entropy term can be computed using, for example, a KDE-based approach, which suffers from the same challenges as in UMI. Alternatively, to bypass estimating P_{XY} at every point, the differential entropy estimation can be done via k-Nearest-Neighbor (kNN) methods (pioneering work in [10]). This KL entropy estimator provides the first step in designing the UMI estimator. However, taking the route of estimating the mutual information via estimating the three differential entropies (two marginal entropies and one joint entropy), it is entirely unclear how to estimate two of these quantities (differential entropy of Y and that of (U,Y)) directly from samples.

Perhaps surprisingly, an innovative approach undertaken in [9] to improve upon three applications of KL estimators provides a solution. The KSG estimator of [9] is based on kNN distance $\rho_{k,i}$ defined as the distance to the k-th nearest-neighbor from (X_i, Y_i) in ℓ_∞ distance, i.e. $\rho_{k,i} = \max\{\|X_{j_k} - X_i\|_\infty, \|Y_{j_k} - Y_i\|_\infty\}$ where (X_{j_k}, Y_{j_k}) is the k-th nearest-neighbor to (X_i, Y_i). Precisely, the KSG estimator is

$$\widehat{I}(X;Y) = \frac{1}{N}\sum_{i=1}^{N}\big(\psi(k) + \psi(N) - \psi(n_{x,i}) - \psi(n_{y,i})\big)\,, \tag{3.1}$$

where $\psi(x)$ is the digamma function, $\psi(x) = \Gamma'(x)/\Gamma(x)$ (for large x, $\psi(x) \approx \log x - 1/(2x)$), and the kNN statistics $n_{x,i}$ and $n_{y,i}$ are defined as

$$n_{x,i} \equiv \sum_{j\neq i}\mathbb{I}\{\|X_j - X_i\|_\infty < \rho_{k,i}\}, \tag{3.2}$$

$$n_{y,i} \equiv \sum_{j\neq i}\mathbb{I}\{\|Y_j - Y_i\|_\infty < \rho_{k,i}\}. \tag{3.3}$$

Note that the number of nearest-neighbors in X and Y are computed with respect to $\rho_{k,i}$ in the joint space (X,Y). This innovative idea not only gives a simple estimator, but also has an advantage of canceling correlations in three entropy estimates, giving an improved performance. However, despite its popularity in practice due to its simplicity, no convergence result has been known until very recently (when [24] showed some consistency and rate of convergence properties).

Inspired by the innovative mutual information estimator in (3.1), we combine importance sampling techniques to adjust for the uniform prior for UMI, and propose a novel estimator. On top of the provable convergence, our estimator has only one hyper-parameter k (besides the choice of bandwidth h_N for estimating the marginal distribution P_X which is a significantly simpler task compared to estimating the joint), which is the number of nearest-neighbors to consider; in practice k is set to an integer such as 4 or 5.

Continuous $\mathcal{X}$. We propose a novel UMI estimator based on the Kraskov mutual information estimator. For a conditional probability density $f_{Y|X}$, we want to compute the uniform mutual information from N i.i.d. samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ that are generated from $f_{Y|X} f_X$ for some prior on X. Our UMI estimator is based on k nearest-neighbor (kNN) statistics. Given a choice of $k \in \mathbb{Z}^+$ and N samples,

$$\widehat{\mathrm{UMI}} \equiv \frac{1}{N} \sum_{i=1}^{N} w_i \left(\psi(k) + \log \frac{N c_{d_x} c_{d_y}}{c_{d_x+d_y}\, n_{x,i}\, n_{y,i}} \right), \tag{3.4}$$

where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, $c_d = \pi^{\frac{d}{2}} / \Gamma(\frac{d}{2}+1)$ is the volume of a d-dimensional unit ball, and w_i is the *self-normalized* importance sampling estimate [25] of $\frac{u(X_i)}{f(X_i)}$:

$$w_i \;\; \equiv \;\; \frac{N/\tilde{f}(X_i)}{\sum_{j=1}^{N} \left(1/\tilde{f}(X_j)\right)} \;, \tag{3.5}$$

where $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is the estimate of $f_X(x)$. We use the standard kernel density estimator with a bandwidth h_N:

$$\tilde{f}(x) \;\; \equiv \;\; \frac{1}{N h_N^{d_x}} \sum_{i=1}^{N} K\Big(\frac{X_i - x}{h_N}\Big) \;. \tag{3.6}$$

We define the kNN statistics $n_{x,i}$ and $n_{y,i}$ as follows. For each sample (X_i, Y_i), calculate the Euclidean distance $\rho_{k,i}$ (as opposed to the ℓ_∞ distance proposed by [9]) to the k-th nearest-neighbor. This determines the (random) number of samples within $\rho_{k,i}$ in $\mathcal{X}$: first $n_{x,i}$ is defined as the same as in (3.2), but with Euclidean distance; second we have a *weighted* number of samples within $\rho_{k,i}$ in $\mathcal{Y}$ as

$$n_{y,i} \equiv \sum_{j \neq i} w_j \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \tag{3.7}$$

Compared to (3.1), we first exchange log function for the digamma functions of N, $n_{x,i}$, and $n_{y,i}$. This step (especially for $n_{x,i}$, and $n_{y,i}$) is crucial for proving convergence. We use ideas from importance sampling and introduce new variables w_i's that capture the correction for the mismatch in the prior. The constants c_{d_x}, c_{d_y}, and $c_{d_x+d_y}$ correct for the volume measured in ℓ_2.

Discrete $\mathcal{X}$. Similarly, for a discrete random variable X, the joint probability density function is denoted by $f(x,y) = p_X(x) f_{Y|X}(y|x)$. We propose a UMI estimator, and overload the same notation for this discrete case.

$$\widehat{\mathrm{UMI}} \equiv \frac{1}{N} \sum_{i=1}^{N} w_{X_i} \left(\psi(k) + \log \frac{N}{n_{X_i}\, n_{y_i}} \right), \tag{3.8}$$

where $n_{X_i} = |\{j \in [N] : j \neq i, X_j = X_i\}|$ is the number of samples j such that $X_j = X_i$, w_{X_i} is the self-normalizing estimate of $1/(|\mathcal{X}| p_X(X_i))$ defined as

$$w_x \equiv \frac{N}{|\mathcal{X}| n_x}, \tag{3.9}$$

and $n_{y,i}$ is the weighted kNN statistics defined as follows. For each sample (X_i, Y_i), let the distance to the k-th nearest-neighbor be $\rho_{k,i}$, where those samples that have the same X value as X_i are considered and the Euclidean distance is measured in $\mathcal{Y}$. We define the *weighted* number of samples within $\rho_{k,i}$ in $\mathcal{Y}$ as

$$n_{y,i} \equiv \sum_{j \neq i} w_{X_j} \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \tag{3.10}$$

3.2 Capacitated Mutual Information

Given standard estimators for mutual information and entropy, it is not at all clear how to *directly* estimate CMI where f_X is changed to the (unknown) optimal input distribution. However, combining the mutual information estimator in (3.1) with importance sampling techniques, we design a novel estimator as a solution to an optimization over the space of the weights. Our estimator has only one hyper-parameter k, the number of nearest-neighbors to consider.

Continuous $\mathcal{X}$. For a conditional distribution $f_{Y|X}$, we compute an estimate of CMI from i.i.d. samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ generated from $f_{Y|X} f_X$ for some prior on X. We introduce a novel CMI estimator that is based on our UMI estimator. Given a choice of $k \in \mathbb{Z}^+$ and N samples, the estimated CMI is the solution of the following constrained optimization:

$$\widehat{\text{CMI}} = \max_{w \in T_{a,N}} \frac{1}{N} \sum_{i=1}^{N} w_i \left(\psi(k) + \log\left(\frac{N c_{d_x} c_{d_y}}{c_{d_x+d_y} n_{x,i} n_{y,i}} \right) \right) , \tag{3.11}$$

where d_x, d_y, $n_{x,i}$, $n_{y,i}$ and c_d are defined in the same way as in (3.4). We optimize over $w_1, \ldots, w_N$ under the second moment constraint, i.e.

$$T_{a,N} = \{ w \in \mathbb{R}^N \,\big|\, w_i \geq 0, \forall i, \frac{1}{N} \sum_{i=1}^{N} w_i = 1, \frac{1}{N} \sum_{i=1}^{N} w_i \|X_i\|^2 \leq a^2 \}. \tag{3.12}$$

Observe that no KDE of P_X is needed for CMI estimation, making it particularly simple and robust.

Discrete $\mathcal{X}$. Similarly, we define the CMI estimate $\widehat{\text{CMI}}$ as the solution of the following constrained optimization:

$$\widehat{\text{CMI}} = \max_{w \in T_\Delta} \quad \tfrac{1}{N} \textstyle\sum_{i=1}^{N} w_{X_i} \left(\psi(k) + \log\left(\tfrac{N}{n_{x,i} n_{y,i}} \right) \right), \tag{3.13}$$

where $n_{x,i}$ and $n_{y,i}$ are defined in (3.8). T_Δ is the set of quantized version of an interval $[C_1, C_2]$ with step size Δ, i.e.

$$\begin{aligned} T_\Delta = \{ w \in \{C_1 + m_i \Delta\}^{|\mathcal{X}|} \,\big|\, & \frac{1}{N} \sum_{x=1}^{|\mathcal{X}|} w_x \in [1 - |\mathcal{X}|\Delta, 1 + |\mathcal{X}|\Delta] \\ & \text{and } m_i \in \{0, 1, \ldots, \lceil (C_2 - C_1)/C_1 \rceil \}, \forall i \, \}. \end{aligned} \tag{3.14}$$

CHAPTER 4

CONVERGENCE GUARANTEES

We show that both the proposed UMI and CMI estimators are consistent under typical assumptions on the distribution. While consistency of estimators in the large sample limit is generally only a (basic) first step in understanding their properties, this is not so for fixed-k nearest-neighbor-based estimators. As far as we know, the only estimator based on fixed-k nearest-neighbors that is known to be consistent is the entropy estimator of [10], and the convergence rate is only known for the univariate case [26] (and that too is under significant assumptions on the univariate density). Our result below for the consistency of the UMI estimator for a discrete alphabet marks another instance where consistency of fixed-k nearest-neighbor-based estimators is established.

4.1 Uniform Mutual Information

As our estimators use the off-the-shelf kernel density estimator of P_X [27, 28] and also the ideas from the nearest-neighbor methods [10], we make assumptions on the conditional density $f_{Y|X}$ that are typical in this literature. One extra assumption we make for UMI is that the Radon-Nikodym derivative $\frac{dP_U}{dP_X}$ is uniformly bounded over the support. This is necessary for controlling the importance-sampling estimates of w_i's. We refer to Assumption 1 in Appendix A for a precise description.

Theorem 1. *Under the Assumption 1 in Appendix A, the UMI estimator converges to the true value in probability, i.e. for all* $\varepsilon > 0$ *and all* $\delta > 0$*,*

$$\lim_{N\to\infty} \mathbb{P}\Big(\big|\widehat{\mathrm{UMI}} - \mathrm{UMI}(f_{Y|X})\big| > \varepsilon \Big) = 0 \ , \tag{4.1}$$

if $k > \max\{d_y/d_x, d_x/d_y\}$ *for continuous* $\mathcal{X}$ *and* $(\log N)^{(1+\delta)d_y} < k < \frac{\sqrt{N}}{5\log N}$

for discrete $\mathcal{X}$.

In practice, we regularize the kNN distance $\rho_{k,i}$ in case it is much smaller than the expected distance of order $N^{-1/(d_x+d_y)}$. For continuous $\mathcal{X}$, we require k to be larger than the ratio of the dimensions, which is a finite constant. For discrete $\mathcal{X}$, however, the effective dimension of $\mathcal{X}$ is zero, which makes the ratio d_y/d_x unbounded. Hence, for the concentration of measure to hold, we need k^{1/d_y} scaling at least logarithmically in the number of samples N.

4.2 Capacitated Mutual Information

We make analogous assumptions which are described precisely in Assumption 2 in Appendix B. The following theorem establishes consistency of our estimator when $\mathcal{X}$ is discrete and we quantize $\mathcal{Y}$. Our analysis needs uniform convergence over all possible choices of the weights w, making the quantization step inevitable; improvements on this technical condition are natural future steps.

Theorem 2. *Under the Assumption 2 in Appendix B, the CMI estimator converges in probability to the true value up to the resolution of the quantization, i.e. if $k > (\log N)^{(1+\delta)d_y}$ for some $\delta > 0$, and $k < \sqrt{N}/(5\log N)$, for all $\varepsilon > 0$ and $\Delta > 0$ and $s(\Delta) = O(\Delta)$,*

$$\lim_{N\to\infty} \mathbb{P}\big(\,\big|\widehat{\mathrm{CMI}} - \mathrm{CMI}(f_{Y|X})\big| > \varepsilon + s(\Delta)\,\big) = 0.$$

CHAPTER 5

NUMERICAL EXPERIMENTS

In this chapter, we provide results for several numerical experiments for UMI and CMI. Experiments include both synthetic simulation and application to the gene pathway influence detection problem.

5.1 Gene Causal Strength from Single Cell Data

We briefly describe the setup of [8] to describe our numerical experiments. Consider a simple genetic pathway: a cascade of genes having expression values X, Y, Z which interact linearly, i.e., $X \to Y \to Z$. A key question of interest in this case is how the signaling in the pathway varies in different conditions of intervention. Let T denote the time after the intervention (for example, after giving a certain drug). Then we may want to compare the strength of the causal relationship between two genes at different times after the intervention. In the experiments, usually samples are taken at very few time points, so T has very small cardinality (for example, before the drug, 10 minutes after the drug and 50 minutes after the drug), but at each given time point, many cells are interrogated so we get samples from the distribution $P_{X,Y,Z;T=t} = P(Y|X;T=t)P(Z|Y;T=t)$. For each value of $T = t$, we observe N_t i.i.d. samples (X_i, Y_i, Z_i), for $i = 1, 2, ..., N_t$ sampled from $P_{X,Y,Z;T=t}$. These samples are obtained using a technique called single-cell mass flow cytometry, see [8] for details. We are interested in obtaining a causal measure $\mathcal{C}(X \to Y; T = t) = \mathcal{C}(P(Y|X;T=t))$ and another measure $\mathcal{C}(Y \to Z; T = t) = \mathcal{C}(P(Z|Y;T=t))$ for each time point t. This measure serves as a high-level summary of how signaling proceeds in the cascade as a function of time, and lets one compare the strengths of a given causal relationship at different points after intervention.

If the drug indeed activates the causal pathway, one may expect the causal relationship to follow a certain *trend*, i.e., at earlier t, the strength of $\mathcal{C}(X \to Y; T = t)$ will be high and at a later value of t, the strength of $\mathcal{C}(Y \to Z; T = t)$ will be high before the effect of the drug wears off, at which time we expect all the relationships to fall back to its low nominal value. Such an analysis is conducted in [8] where the causal strength function $\mathcal{C}$ is evaluated via the so-called DREMI estimator (essentially a version of UMI estimation with a "channel amplification" step and careful choice of hyper parameters therein — no theoretical properties of this estimator were evaluated). In that paper, it is shown that, for two example pathways, DREMI recovers the correct trend, i.e., it correctly identifies the time at which each causal relationship is expected to peak as per prior biological knowledge. This demonstrates the utility of DREMI for causal strength inference in gene networks (see Figure 6 of [8]). The authors there also demonstrate that other metrics which depend on the whole joint distribution, such as mutual information, maximal information coefficient, and correlation do not capture the trend. As an aside, we note that a somewhat different set of "trend spotting" estimators, primarily trying to find genes which demonstrate a monotonic trend over time from single-cell RNA-sequencing data, have been proposed very recently in [29].

In the thesis, we have studied influence measures axiomatically and proposed the UMI and CMI measures. It is natural to apply our estimators to *each time point* in the same setting as [8] — and look to understand two distinct issues in our experiments with the flow-cytometry data. The first is whether the proposed quantities of UMI and CMI are able to capture the same biological trend that DREMI was able to capture. The second question relates to the sample complexity: How does the ability to recover the trend vary as a function of the sample complexity? To study this, we subsample the original data from [8] multiple times (100 in the experiments) at each subsampling ratio and compute the fraction of times we recover the true biological trend. This is plotted in Figure 5.1. The figure demonstrates that when the whole dataset is made available, UMI and CMI are able to spot the trend correctly (just as DREMI does). When fewer samples are available, UMI uniformly dominates DREMI and, in turn, CMI uniformly dominates UMI in terms of capturing the biological trend as a function of number of samples available. We believe that this strong empirical evidence lends credence to our approach. For completeness, we note that the datasets represented in

Figure 5.1 refer to regular T-cells (top figure) and T-cells exposed with an antigen (bottom figure), for which we expect different biological trends, but both of which are correctly captured by our metrics.

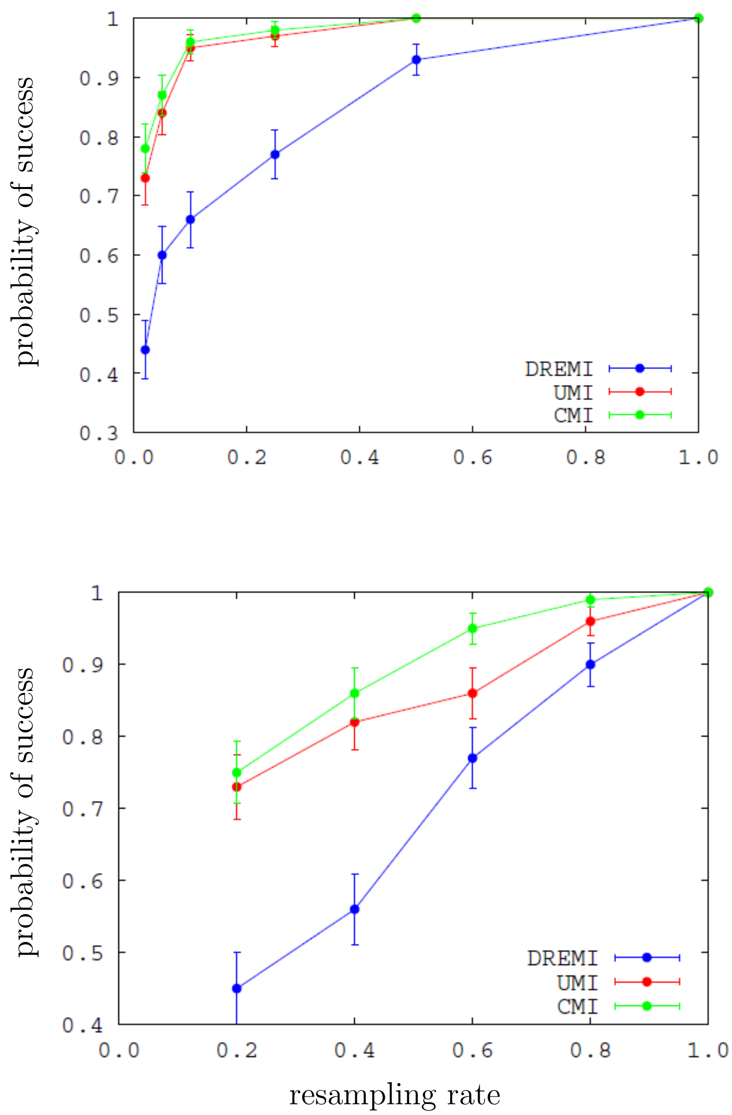

Figure 5.1: CMI and UMI estimators significantly improve over DREMI in capturing the biological trend in flow-cytometry data: the figures above refer to the same setting as Figure 6 of [8].

5.2 Synthetic Data

We demonstrate the accuracy of the proposed UMI and CMI estimators on synthetic experiments. We generate N samples from P_{XY} where X is distributed as a beta distribution Beta$(1.5, 1.5)$ and $Y = X + N$, $N \sim \mathcal{N}(0, \sigma^2)$, independent of X. We present three results with varying $\sigma^2 \in \{0.09, 0.36, 1.0\}$. Figure 5.2 shows the estimate of UMI, averaged over 100 instances. This is compared to the ground truth and the state-of-the-art partition-based estimators from [11]. The ground truth has been computed via simulations with 8192 samples from the desired distribution $P_{Y|X}U_X$ using Kraskov's mutual information estimator [9].

For CMI, we use exactly the same distribution P_{XY} as in UMI, but with varying $\sigma^2 \in \{0.36, 1.0, 2.25\}$, which is illustrated in Figure 5.3. Under the power constraint, the ground truth is given by $\frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_N^2}) = \frac{1}{2}\log(1 + 1/16\sigma^2)$. The proposed CMI estimator is compared agains the Blahut-Arimoto algorithm [30, 31] for computing discrete channel capacity, applied to quantized data. Both figures illustrate that the proposed estimators significantly improve over the state-of-the-art partition-based methods, in terms of sample complexity.

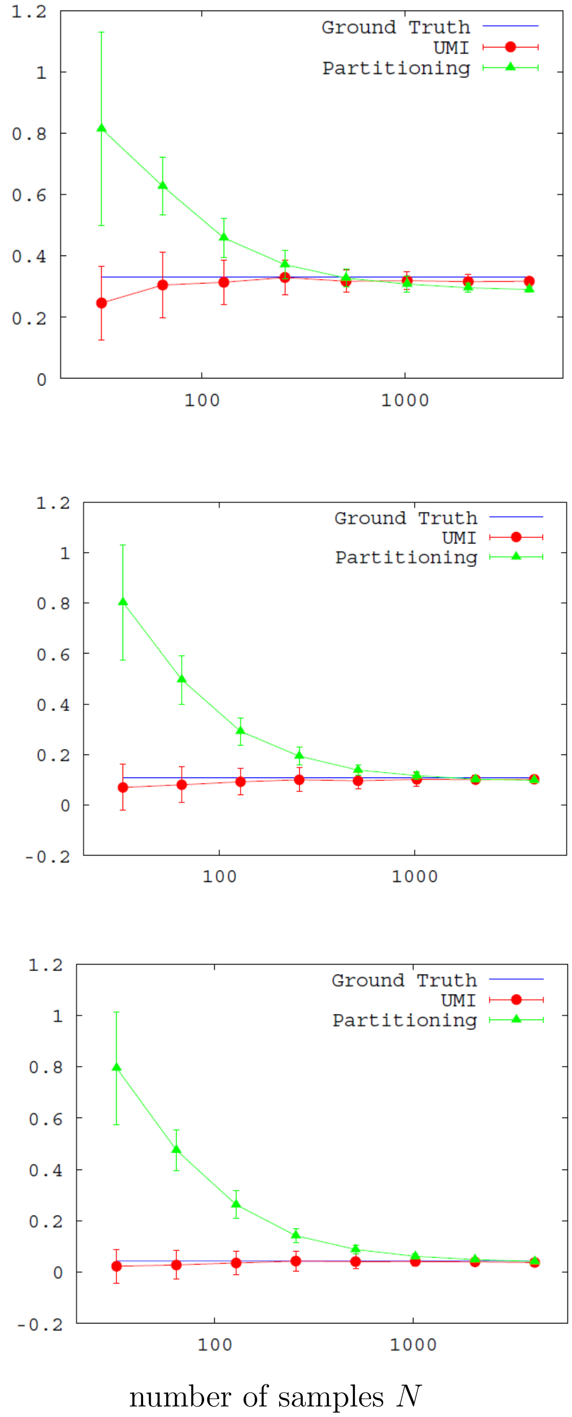

Figure 5.2: The proposed UMI estimator significantly outperforms partition-based methods [11] in sample complexity. Additive Gaussian channels are used with varying variances σ^2: 0.09 (top), 0.36 (middle), and 1.0 (bottom).

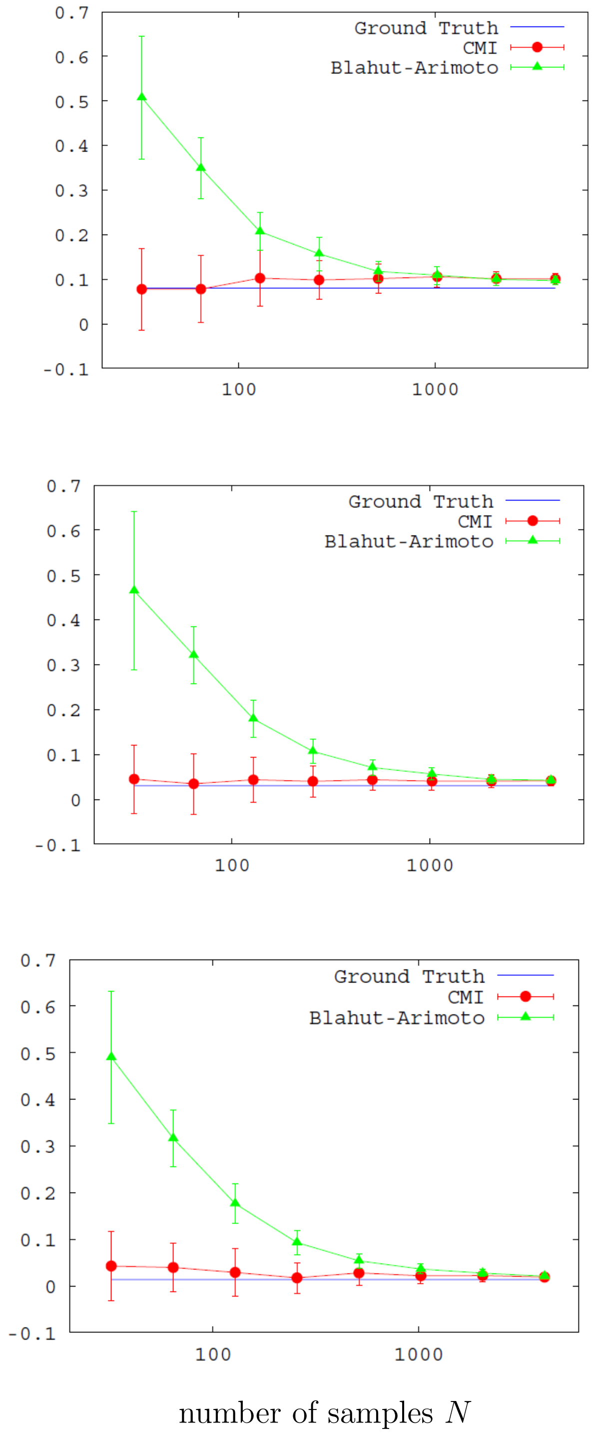

Figure 5.3: The proposed CMI estimator significantly outperforms partition-based methods [30, 31] in sample complexity. Additive Gaussian channels are used with varying variances σ^2: 0.36 (top), 1.0 (middle), and 2.25 (bottom).

CHAPTER 6

CONCLUSION AND DISCUSSION

In the thesis we have proposed novel information theoretic measures of potential influence of one variable on another, as well as provided novel estimators to compute the measures from i.i.d. samples. The technical innovation has been in proposing these estimators, by combining separate threads of methods in statistics (including importance sampling and nearest-neighbor methods). The consistency proofs suggest that a similar analysis the very popular estimator of (traditional) mutual information in [9] can be conducted successfully; such work has been recently conducted in [24]. Several other issues in statistical estimation theory intersect with our current work and we discuss some of these topics below.

- The main technical results of the thesis have been weak consistency of the proposed estimators. Proving stronger consistency guarantees and rates of convergence would be natural improvements, albeit challenging ones. Rates of convergence in the nearest-neighbor methods are barely known in the literature even for traditional information theoretic quantities: for instance, [26] derives a $\sqrt{N}$ consistency for the single-dimensional case of differential entropy estimation (under strong assumptions on the underlying pdf), leaving higher-dimensional scenarios open, and which recently have been successfully addressed in [24].

- There is a natural generalization of our estimators when the alphabet $\mathcal{Y}$ is high-dimensional, using the kNN approach (just as in the differential entropy estimator of [10] or in the mutual information estimator of [9]). However, very recent works [20, 21, 32] have shown that boundary biases common in high-dimensional scenarios are much better handled using local parametric methods (as in [33, 34]). Adapting these approaches to the estimators for UMI and CMI is an interesting direction of future research.

- We have considered both the case of discrete and (single-dimensional) continuous alphabet $\mathcal{X}$. The scenario of high-dimensional $\mathcal{X}$ is significantly more challenging for CMI estimation: this is because of the (vastly) expanded space of distributions over which the optimization can be performed. Also challenging is to consider application-specific regularization of the inputs in this scenario.

- While the focus of the thesis has been on quantifying potential causal influence, a related question involves testing the *direction* of causality for a pair of random variables. This is a widely studied topic with a long lineage [4] but also of strong topical interest [7, 35, 6, 36]. A natural inclination is to explore the efficacy of UMI and CMI measures to test for direction of causality — especially in the context of the benchmark data sets collected in [6]. Our results are as follows: UMI has a 45% probability to predict the correct direction. CMI gives 53% probability. Directly comparing the marginal entropy $H(X)$ and $H(Y)$ by the estimator in [10] also only provides 45% accuracy. While in [6], different entropy estimators (with appropriate hyper parameter choices) were applied to get an accuracy up to 60%-70%. Further research is needed to shed conclusive light, although we point out that the benchmark datasets in [6] have substantial confounding factors that make causal direction hard to measure in the first place.

- The axiomatic derivation of potential causal influence naturally suggests CMI as an appropriate measure. We are also able to show that a more general quantity, the so-called Rényi capacity, also meets the axioms. For any $\lambda > 0, \lambda \neq 1$, define Rényi entropy:

$$H_\lambda(P) := \frac{1}{\lambda - 1} \log \mathbb{E}_P[(dP)^\lambda] \tag{6.1}$$

and Rényi divergence:

$$D_\lambda(P\|Q) := \frac{1}{\lambda - 1} \log \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^\lambda\right]. \tag{6.2}$$

Now define the *asymmetric* information measure [37]:

$$K_\lambda(P_X P_{Y|X}) := \inf_{Q_Y} D_\lambda(P_{XY} \| P_X Q_Y), \tag{6.3}$$

which converges to the traditional mutual information when $\lambda \to 1$. Now we can define the Rényi capacity for any parameter λ and any fixed conditional distribution $P_{Y|X}$:

$$\mathrm{CMI}_\lambda := \sup_{P_X} \inf_{Q_Y} D_\lambda(P_{XY} \| P_X Q_Y). \tag{6.4}$$

Observe that as $\lambda \to 1$, we have $\mathrm{CMI}_\lambda \to \mathrm{CMI}$, the traditional Shannon capacity. We observe the following.

Proposition 2. *For any $\lambda > 0, \lambda \neq 1$ we have that CMI_λ satisfies the axioms in Chapter 2.*

The proof is available in Appendix C. In the light of this result, it would be interesting to design estimators for the more general family of Rényi capacity measures and confirm their performance on empirical tasks such as the ones studied in [8]. It would also be very interesting to understand the role of additional axioms that would lead to the uniqueness of Shannon capacity (in the same spirit as entropy being uniquely characterized by somewhat similar axioms [3]).

- Another measure of information commonly used to quantify causal strength is *directed* mutual information (DMI): for a pair of random *vectors* $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ this quantity is defined as [38]

$$\sum_{k=1}^{n} I(X_1, \ldots, X_k; Y_k | Y_1, \ldots, Y_{k-1}). \tag{6.5}$$

While this quantity has been posited as a relevant metric in studying causality for a pair of sequences of observations (typically a time series) [39, 40], in the context of the thesis it simply reduces to the traditional mutual information $I(X;Y)$ (since we are studying only single random variables X and Y). If the random variables X and Y are vectors (sequences "in time") and a causal structure is suspected "within" the components of X and Y, then CMI specializes to this setting readily and could be used to quantify the causal influence, by computing the Shannon capacity of the time varying (Markov) channel from X and Y. In this context, CMI differs from DMI in the same sense as in the scalar random variable originally introduced in the thesis: CMI captures *potential* influence while DMI measures *factual* influence.

- Finally, a comment on the optimization problem in CMI estimation: the optimization problem involving the w_i's is not necessarily a concave program for a given sample realization, although this program converges to that of Shannon capacity computation (involving maximizing mutual information), which is a concave function of the input probability distribution. Standard (stochastic) gradient decent is used in our experiments, and we did not face any disparity in convergent values over the set of synthetic experiments we conducted.

APPENDIX A

PROOF OF THEOREM 1

We present the proof of the theorem for two separate UMI estimators: first for continuous X and next for discrete X. We first state the formal assumptions under which the theorem holds.

Assumption 1. *For continuous $\mathcal{X}$, define*

$$f_U(y) \equiv \int_x u(x) f_{Y|X}(y|x)dx, \tag{A.1}$$

$$f_U(x\big|y) \equiv \frac{u(x)f_{Y|X}(y|x)}{f_U(y)}. \tag{A.2}$$

We make the following assumptions:

(*a*) $\int f_{X,Y}(x,y)\big(\log f(x,y)\big)^2 dxdy < \infty$.

(*b*) *There exists a finite constant C such that the Hessian matrix of $H(f)$ and $H(f_U)$ exists and $\max\{\|H(f)\|_2, \|H(f_U)\|_2\} < C$ almost everywhere.*

(*c*) *There exists a positive constant C' such that the conditional pdfs satisfy $f_{Y|X}(y\big|x) < C'$ and $f_U(x\big|y) < C'$ almost everywhere.*

(*d*) *There exist positive constants $C_1 < C_2$ such that the marginal pdf satisfy, almost everywhere,*

$$\frac{C_1}{\mu(\mathcal{X})} < f_X(x) < \frac{C_2}{\mu(\mathcal{X})} .$$

(*e*) *The bandwidth h_N of the kernel density estimator is chosen as $h_N = \frac{1}{2}N^{-1/(2d_x+3)}$.*

For discrete $\mathcal{X}$, define

$$f_u(y) \equiv \sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|} f_{Y|X}(y|x). \tag{A.3}$$

We make the following assumptions:

(a) $\int f_{Y|X}(y|x)\big(\log f_{Y|X}(y|x)\big)^2 dy < \infty$, *for all* $x \in \mathcal{X}$.

(b) *There exists a finite constant* C *such that the Hessian matrix of* $H(f_{Y|X})$ *exists and* $\|H(f_{Y|X})\|_2 < C$ *almost everywhere, for all* $x \in \mathcal{X}$.

(c) *There exists a finite constant* C' *such that the conditional pdf* $f_{Y|X}(y\big|x) < C'$ *almost everywhere, for all* $x \in \mathcal{X}$.

(d) *There exists finite constants* $C_1 < C_2$ *such that the prior* $p_X(x) > C_1/|\mathcal{X}|$ *and* $f_X(x) < C_2/|\mathcal{X}|$ *almost everywhere.*

A.1 The Continuous Alphabet Case

Given these assumptions, we define

$$g(X_i, Y_i) \equiv \psi(k) + \log(N) + \log\left(\frac{c_{d_x} c_{d_y}}{c_{d_x+d_y}}\right) - \big(\log(n_{x,i}) + \log(n_{y,i})\big)\,, \tag{A.4}$$

such that $\hat{I}^U_{k,N}(X,Y) = \frac{1}{N}\sum_{i=1}^N w_i\, g(X_i, Y_i)$. Define each quantity with the true prior $f_X(x)$ as

$$w'_i \equiv \frac{u(X_i)}{f_X(X_i)}\,, \tag{A.5}$$

$$n'_{y,i} \equiv \sum_{j\neq i} w'_j \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}\,, \tag{A.6}$$

$$g'(X_i, Y_i) \equiv \psi(k) + \log(N) + \log\left(\frac{c_{d_x} c_{d_y}}{c_{d_x+d_y}}\right) - \big(\log(n_{x,i}) + \log(n'_{y,i})\big)\,, \tag{A.7}$$

With U_X equal to the uniform distribution on the support of X, we apply the triangle inequality to show that each term converges to zero in probability.

$$\begin{aligned}
& \left|\hat{I}^U_{k,N}(X,Y) - I^U(f_{Y|X})\right| \\
= & \left|\frac{1}{N}\sum_{i=1}^{N} w_i g(X_i,Y_i) - \iint u(x) f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{\int u(x') f_{Y|X}(y|x')dx'} dxdy\right| \\
\leq & \frac{1}{N}\left|\sum_{i=1}^{N}\left(w_i g(X_i,Y_i) - w'_i g'(X_i,Y_i)\right)\right| \qquad \text{(A.8)} \\
+ & \left|\frac{1}{N}\sum_{i=1}^{N} w'_i g'(X_i,Y_i) - \iint u(x) f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{\int u(x') f_{Y|X}(y|x')dx'} dxdy\right|, \qquad \text{(A.9)}
\end{aligned}$$

The first term (A.8) captures the error in the kernel density estimator and we have the following claim, whose proof is delegated to Appendix A.1.1.

Lemma 1. *The term in Equation* (A.8) *converges to 0 as $N \to \infty$ in probability.*

The second term in the error (A.9) comes from the sample noise in density estimation. Similar to the decomposition of mutual information, $I(X;Y) = H(X) + H(Y) - H(X,Y)$, we decompose our estimator into three terms:

$$\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{N} w'_i g'(X_i,Y_i) \\
= & \widehat{H}^U_{k,N}(X) + \widehat{H}^U_{k,N}(Y) - \widehat{H}^U_{k,N}(X,Y) \\
& - \sum_{i=1}^{N} \frac{w'_i}{N}\left(2\log(N-1) - \psi(N) - \log(N)\right), \qquad \text{(A.10)}
\end{aligned}$$

where

$$\widehat{H}^U_{k,N}(X,Y) \equiv \sum_{i=1}^N \frac{w'_i}{N}\big(-\psi(k)+\psi(N)+\log c_{d_x+d_y}+(d_x+d_y)\log\rho_{k,i}\big) , \tag{A.11}$$

$$\widehat{H}^U_{k,N}(X) \equiv \sum_{i=1}^N \frac{w'_i}{N}\big(-\log n_{x,i}+\log(N-1)+\log c_{d_x}+d_x\log\rho_{k,i}\big) , \tag{A.12}$$

$$\widehat{H}^U_{k,N}(Y) \equiv \sum_{i=1}^N \frac{w'_i}{N}\big(-\log n_{y,i}+\log(N-1)+\log c_{d_y}+d_y\log\rho_{k,i}\big) . \tag{A.13}$$

Notice that$\sum_{i=1}^N \frac{w'_i}{N}(2\log(N-1)-\psi(N)-\log(N))$ goes to 0 as N goes to infinity. The desired claim follows directly from the following two lemmas showing the convergence each entropy estimates to corresponding entropies under UMI.

Lemma 2. *Under the hypotheses of Theorem 1, for all* $\varepsilon > 0$

$$\lim_{N\to\infty}\mathbb{P}\Big(\Big|\iint u(x)f_{Y|X}(y|x)\log(f_{Y|X}(y|x)u(x))dxdy + \widehat{H}^U_{k,N}(X,Y)\Big| > \varepsilon\Big) = 0 . \tag{A.14}$$

Lemma 3. *Under the hypotheses of Theorem 1, for all* $\varepsilon > 0$

$$\lim_{N\to\infty}\mathbb{P}\Big(\Big|\iint u(x)f_{Y|X}(y|x)\big(\log f_X(x)+\log f_U(y)\big)dxdy + \widehat{H}^U_{k,N}(X)+\widehat{H}^U_{k,N}(Y)\Big| > \varepsilon\Big) = 0 , \tag{A.15}$$

where $f_U(y) = \int f_{Y|X}(y|x)u(x)dx$.

A crucial technical idea in proving these lemmas is the concept of importance sampling. For any function $h(X_i, Y_i)$, the *importance sampling* estimate of $\mathbb{E}[h]$ is given by

$$\tilde{h}_N = \frac{1}{N}\sum_{i=1}^N w'_i h(X_i, Y_i) , \tag{A.16}$$

where $w'_i = Nu(X_i)/f(X_i)$. The following lemma gives the almost sure convergence of $\tilde{h}_n$.

Lemma 4 (Theorem 9.1 in [41])**.** *Assume* $\mathbb{E}[h] = \iint u(x) f_{Y|X}(y|x) h(x,y) dx dy$ *exists, then*

$$\mathbb{P}\big(\lim_{n\to\infty} \tilde{h}_n = \mathbb{E}[h] \big) = 1 \ .$$

A.1.1 Proof of Lemma 1

The term in Equation (A.8) is upper bounded by:

$$\begin{aligned}
& \frac{1}{N}\Big| \sum_{i=1}^{N} \big(w_i g(X_i, Y_i) - w_i' g'(X_i, Y_i) \big) \Big| \\
\leq \ & \frac{1}{N} \sum_{i=1}^{N} \big| w_i g(X_i, Y_i) - w_i' g'(X_i, Y_i) \big| \\
\leq \ & \frac{1}{N} \sum_{i=1}^{N} \Big(|w_i - w_i'| \, |g'(X_i, Y_i)| + w_i \, \big| g(X_i, Y_i) - g'(X_i, Y_i) \big| \Big) \\
= \ & \frac{1}{N} \sum_{i=1}^{N} \Big(|w_i - w_i'| \, |g'(X_i, Y_i)| + w_i \, \big| \log(n_{y,i}) - \log(n_{y,i}') \big| \Big) \\
\leq \ & \frac{1}{N} \sum_{i=1}^{N} \Big(|w_i - w_i'| \, |g'(X_i, Y_i)| + w_i \, \big| n_{y,i} - n_{y,i}' \big| \big(\frac{1}{2n_{y,i}} + \frac{1}{2n_{y,i}'} \big) \Big) \\
\leq \ & \frac{1}{N} \sum_{i=1}^{N} |w_i - w_i'| \, |g'(X_i, Y_i)| + \sum_{i=1}^{N} \frac{w_i}{N} \Big(\frac{\sum_{j\neq i} \mathbb{I}\{\|Y_j - Y_i| < \rho_{k,i}\} |w_j' - w_j|}{2\sum_{j\neq i} \mathbb{I}\{\|Y_j - Y_i| < \rho_{k,i}\} w_j'} \\
& + \frac{\sum_{j\neq i} \mathbb{I}\{\|Y_j - Y_i| < \rho_{k,i}\} |w_j' - w_j|}{2\sum_{j\neq i} \mathbb{I}\{\|Y_j - Y_i| < \rho_{k,i}\} w_j} \Big) \\
\leq \ & \frac{1}{N} \sum_{i=1}^{N} |w_i - w_i'| \, |g'(X_i, Y_i)| + \sum_{i=1}^{N} \frac{w_i}{N} \big(\frac{\max_{1\leq j\leq N} |w_j' - w_j|}{2\min_{1\leq j\leq N} w_j'} \\
& + \frac{\max_{1\leq j\leq N} |w_j' - w_j|}{2\min_{1\leq j\leq N} w_j'} \big) \\
\leq \ & \max_{1\leq i\leq N} |w_i - w_i'| \big(\max_{1\leq i\leq N} |g'(X_i, Y_i)| + \frac{1}{2\min_{1\leq j\leq N} w_j'} + \frac{1}{2\min_{1\leq j\leq N} w_j} \big) \ ,
\end{aligned} \tag{A.17}$$

where the last inequality follows from the fact that $\sum_{i=1}^{N} w_i = N$. We upper bound each term as follows.

$$w'_j \;=\; \frac{u(X_i)}{f_X(X_i)} \geq \frac{1/\mu(K)}{C_2/\mu(K)} = 1/C_2. \tag{A.18}$$

Similarly we have $w'_j \leq 1/C_1$. This implies that $n'_{y,i} = \sum_{j\neq i} w'_j \mathbb{I}\{\|Y_i - Y_j\| \leq \rho_{k,i}\} \geq k/C_2$. For finite k and sufficiently large N, we have:

$$\begin{aligned} g'(X_i,Y_i) &= \psi(k) + \log(N) + \log(\frac{c_{d_x}c_{d_y}}{c_{d_x+d_y}}) - \big(\log(n_{x,i}) + \log(n_{y,i})\big) \\ &\leq \psi(k) + \log(N) + \log(\frac{c_{d_x}c_{d_y}}{c_{d_x+d_y}}) - \big(\log(k) + \log(k/C_2)\big) \\ &\leq 2\log N\ , \end{aligned} \tag{A.19}$$

and similarly, using the fact that $n_{y,i} = \sum_{j\neq i} w'_j \mathbb{I}\{\|Y_i - Y_j\| \leq \rho_{k,i}\} \leq N/C_1$,

$$\begin{aligned} g'(X_i,Y_i) &\geq \psi(k) + \log(N) + \log(\frac{c_{d_x}c_{d_y}}{c_{d_x+d_y}}) - \big(\log(N) + \log(N/C_1)\big) \\ &\geq -2\log N. \end{aligned} \tag{A.20}$$

We claim that for sufficiently large N such that $\log N > \max\{C_2\varepsilon/3, 3C_2/2\}$, if $|w_i - w'_i| < \varepsilon/(3\log N)$ for all i, then (A.17) is upper bounded by ε.

$$\begin{aligned} & \max_{1\leq i\leq N} |w_i - w'_i| \big(\max_{1\leq i\leq N} |g'(X_i,Y_i)| + \frac{1}{2\min_{1\leq j\leq N} w'_j} + \frac{1}{2\min_{1\leq j\leq N} w_j} \big) \\ \leq\; & \frac{\varepsilon}{3\log N}\big(2\log N + \frac{C_2}{2} + \frac{1}{2/C_2 - \frac{\varepsilon}{3\log N}} \big) \\ \leq\; & \frac{\varepsilon}{3\log N}\big(2\log N + \frac{C_2}{2} + C_2 \big) \leq \varepsilon. \end{aligned} \tag{A.21}$$

Putting these bounds together, we have, for any $\varepsilon > 0$ and sufficiently large N,

$$\begin{aligned} & \mathbb{P}\Big(\frac{1}{N} \big| \sum_{i=1}^{N} \big(w_i g(X_i,Y_i) - w'_i g'(X_i,Y_i) \big) \big| > \varepsilon \Big) \\ \leq\; & \mathbb{P}\Big(\max_{1\leq i\leq N} |w_i - w'_i| > \frac{\varepsilon}{3\log N} \Big)\ . \end{aligned} \tag{A.22}$$

Define

$$w_i'' = \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)} \; , \tag{A.23}$$

and applying the triangle inequality and union bound for (A.22), we have

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le i\le N} |w_i - w_i'| > \frac{\varepsilon}{3\log N} \Big) \\
\le\;& \mathbb{P}\Big(\max_{1\le i\le N} |w_i' - w_i''| + \max_{1\le i\le N} |w_i'' - w_i| > \frac{\varepsilon}{3\log N} \Big) \\
\le\;& \mathbb{P}\Big(\max_{1\le i\le N} |w_i' - w_i''| > \frac{\varepsilon}{6\log N} \Big) \qquad \text{(A.24)}\\
&+\mathbb{P}\Big(\max_{1\le i\le N} |w_i'' - w_i| > \frac{\varepsilon}{6\log N} \Big). \qquad \text{(A.25)}
\end{aligned}$$

For (A.24), recall that $w_i' = u(X_i)/f_X(X_i)$. Since $u(X_i)/f_X(X_i) \in [1/C_2, 1/C_1]$ for all i. Therefore,

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le i\le N} |w_i' - w_i''| > \frac{\varepsilon}{6\log N} \Big) \\
=\;& \mathbb{P}\Big(\max_{1\le i\le N} |\frac{u(X_i)}{f_X(X_i)} - \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)}| > \frac{\varepsilon}{6\log N} \Big) \\
=\;& \mathbb{P}\Big(\big(\max_{1\le i\le N} \frac{u(X_i)}{f_X(X_i)} \big| 1 - \frac{N}{\sum_{j=1}^N \big(u(X_j)/f_X(X_j)\big)} \big| \big) > \frac{\varepsilon}{6\log N} \Big) \\
\le\;& \mathbb{P}\Big(\frac{1}{C_1} \Big| \sum_{j=1}^N \frac{u(X_j)}{f_X(X_j)} - N \Big| > \frac{\varepsilon}{6\log N} \sum_{j=1}^N \frac{u(X_j)}{f_X(X_j)} \Big). \\
\le\;& \mathbb{P}\Big(\Big| \sum_{j=1}^N \frac{u(X_j)}{f_X(X_j)} - N \Big| > \frac{\varepsilon}{6\log N} \frac{NC_1}{C_2} \Big). \qquad \text{(A.26)}
\end{aligned}$$

Note that $w_j' = \frac{u(X_j)}{f_X(X_j)}$ are i.i.d. random variables with $\mathbb{E}[w_j'] = \int f_X(x) \frac{u(x)}{f_X(x)} dx = 1$ and $w_j' \in [1/C_2, 1/C_1]$. Therefore, by Hoeffding's inequality, we obtain:

$$\begin{aligned}
& \mathbb{P}\Big(\big| \sum_{j=1}^{N} \frac{u(X_j)}{f_X(X_j)} - N \big| > \frac{\varepsilon N C_1}{6 \log N C_2} \Big) \\
\leq \; & 2 \exp\left\{ - \frac{2 \big(\frac{\varepsilon N C_1}{6 \log N C_2} \big)^2}{N (1/C_1 - 1/C_2)^2} \right\} \\
= \; & 2 \exp\left\{ - \frac{\varepsilon^2 N C_1^4}{18 (\log N)^2 (C_2 - C_1)^2} \right\} , \qquad \text{(A.27)}
\end{aligned}$$

which shows that the probability in (A.24) goes to 0 as N goes to infinity. For the probability in (A.25), recall that for any i,

$$|w_i'' - w_i| \;=\; \Big| \frac{N / f_X(X_i)}{\sum_{j=1}^{N} \big(1/f_X(X_j)\big)} - \frac{N / \tilde{f}_X(X_i)}{\sum_{j=1}^{N} \big(1/\tilde{f}_X(X_j)\big)} \Big| . \qquad \text{(A.28)}$$

We use the following lemma that shows an upper bound for the error of the kernel density estimator.

$$\tilde{f}_X(X_i) = \frac{1}{(N-1) h_N^{d_x}} \sum_{j \neq i} K\Big(\frac{X_j - X_i}{h_N}\Big). \qquad \text{(A.29)}$$

Lemma 5. *Assume that $K(u) \leq A$ for all u, $\kappa_j(K) = \int_{\mathbb{R}^{d_x}} \|u\|^j K(u) du < +\infty$ for any positive integer $j \geq 1$ and $\int_{\mathbb{R}^{d_x}} u K(u) du = 0$. By choosing $h_N = \frac{1}{2} N^{-1/(2d_x+3)}$, we have for a given $i \in \{1, \ldots, N\}$,*

$$\mathbb{P}\Big(\big| \tilde{f}_X(X_i) - f_X(X_i) \big| > N^{-1/(2d_x+3)} \Big) \leq 2 \exp\{ - \frac{N^{1/(2d_x+3)}}{16 A^2} \}. \qquad \text{(A.30)}$$

Applying the union bound we can get that with probability at least $1 - 2N \exp\{-\frac{N^{1/(2d_x+3)}}{16A^2}\}$, we have

$$|\tilde{f}_X(X_i) - f_X(X_i)| \;<\; N^{-1/(2d_x+3)} \qquad \text{(A.31)}$$

for all i. When this bound holds, we claim that the event inside of the probability in (A.25) holds for sufficiently large N. Together with (A.27), this proves the desired claim: for given $\varepsilon > 0$ and large enough N,

$$\frac{1}{N} \Big| \sum_{i=1}^{N} \big(w_i g(X_i, Y_i) - w_i' g'(X_i, Y_i) \big) \Big| \;\leq\; \varepsilon , \qquad \text{(A.32)}$$

with probability at least $1-2N\exp\{-\frac{N^{1/(2d_x+3)}}{16A^2}\}-2\exp\{-\frac{\varepsilon^2 NC_1^4}{18(\log N)^2(C_2-C_1)^2}\}$.

Now, we are left to show that (A.31) implies the event inside of the probability in (A.25). Given (A.31), we have

$$|\frac{\tilde{f}_X(X_i)}{f_X(X_i)}-1| < \frac{N^{-1/(2d_x+3)}}{f_X(X_i)} < \frac{\mu(K)N^{-1/(2d_x+3)}}{C_1}\ , \tag{A.33}$$

for all i. Therefore, for sufficiently large N, $w_i - w_i'$ is lower bounded by

$$\begin{aligned}
w_i - w_i'' &= \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)} - \frac{N/\tilde{f}_X(X_i)}{\sum_{j=1}^N \big(1/\tilde{f}_X(X_j)\big)} \\
&\geq \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)}\big(1-\frac{1+\frac{\mu(K)N^{-1/(2d_x+3)}}{C_1}}{1-\frac{\mu(K)N^{-1/(2d_x+3)}}{C_1}}\big) \\
&\geq \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)}\big(1-(1+\frac{3\mu(K)N^{-1/(2d_x+3)}}{C_1})\big) \qquad \text{(A.34)} \\
&\geq -\frac{3C_2\mu(K)N^{-1/(2d_x+3)}}{C_1^2}\ , \qquad \text{(A.35)}
\end{aligned}$$

where (A.34) follows from the fact that $(1+a)/(1-a) \leq 1+3a$ for $a \in [0, 1/3]$, and (A.35) follows from the fact that $C_1/\mu(K) \leq f_X(x) \leq C_2/\mu(K)$. Similarly, it is upper bounded by

$$\begin{aligned}
w_i - w_i'' &= \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)} - \frac{N/\tilde{f}_X(X_i)}{\sum_{j=1}^N \big(1/\tilde{f}_X(X_j)\big)} \\
&\leq \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)}\big(1-\frac{1-\frac{\mu(K)N^{-1/(2d_x+3)}}{C_1}}{1+\frac{\mu(K)N^{-1/(2d_x+3)}}{C_1}}\big) \\
&\leq \frac{N/f_X(X_i)}{\sum_{j=1}^N \big(1/f_X(X_j)\big)}\big(1-(1-\frac{3\mu(K)N^{-1/(2d_x+3)}}{C_1})\big) \\
&\leq \frac{3C_2\mu(K)N^{-1/(2d_x+3)}}{C_1^2}\ . \qquad \text{(A.36)}
\end{aligned}$$

Here (A.36) comes from the fact that $(1-a)/(1+a) \geq 1-3a$ for all $a \geq 0$. Therefore, $|w_i - w_i''| \leq 3C_2\mu(K)N^{-1/(2d_x+3)}/C_1^2$. For a given $\varepsilon > 0$ and for sufficiently large N such that $\varepsilon/(6\log N) \geq 3C_2\mu(K)N^{-1/(2d_x+3)}/C_1^2$, we have

$$\begin{aligned}
& \mathbb{P}\Big(\max_{1\leq i\leq N} |w_i'' - w_i| > \frac{\varepsilon}{6\log N} \Big) \\
\leq \;& \mathbb{P}\Big(\max_{1\leq i\leq N} |w_i - w_i''| > \frac{3C_2\mu(K)N^{-1/(2d_x+3)}}{C_1^2} \Big) \\
\leq \;& \mathbb{P}\Big(\forall i\ , |\tilde{f}_X(X_i) - f_X(X_i)| < N^{-1/(2d_x+3)} \Big) \ . \qquad \text{(A.37)}
\end{aligned}$$

Together with (A.22) and (A.27), this proves the desired convergence of the first term (A.8).

A.1.2 Proof of Lemma 5

Given $X_i = x$, denote $a_j = \frac{1}{h_N^{d_x}} K(\frac{X_j - x}{h_N})$ such that $\tilde{f}_X(X_i) = \frac{1}{N}\sum_{j=1}^N a_j$. For sufficiently small h_N, the mean of a_j is given by:

$$\begin{aligned}
& \mathbb{E}[a_j] \\
= \;& \int_{z\in\mathbb{R}^{d_X}} \frac{1}{h_N^{d_x}} K(\frac{z - X_i}{h_N}) f_X(z) dz \\
= \;& \int_{u\in\mathbb{R}^{d_X}} K(u) f_X(x + h_N u) du \\
= \;& \int_{u\in\mathbb{R}^{d_X}} K(u)\big(f_X(x) + h_N u^T \nabla f_X(x) + h_N^2 u^T H(f_X)(x) u + o(h_N^2)\big) du \\
= \;& f_X(x) + h_N^2 C \int_{u\in\mathbb{R}^{d_X}} \|u\|^2 K(u) du + o(h_N^2)\ , \qquad \text{(A.38)}
\end{aligned}$$

where we used the fact that the kernel is centered such that $\int K(u) a^T u du = 0$. For sufficiently small h_N, we obtain $|\mathbb{E}[a_j] - f_X(x)| < h_N$. Therefore,

$$\begin{aligned}
& \mathbb{P}\Big(\big|\, \tilde{f}_X(X_i) - f_X(X_i) \,\big| > N^{-1/(2d_x+3)} \Big| X_i = x \Big) \\
\leq \;& \mathbb{P}\big(\big| \sum_{j=1}^N a_j - N f_X(X_i) \big| > N^{(2d_x+2)/(2d_x+3)} \Big| X_i = x \big) \\
\leq \;& \mathbb{P}\big(\big| \sum_{j\neq i} a_j - (N-1)\mathbb{E}[a_j] \big| > N^{(2d_x+2)/(2d_x+3)} \\
& \quad - \big| a_i - N f_X(x) + (N-1)\mathbb{E}[a_j] \big| \Big| X_i = x \big)\ . \qquad \text{(A.39)}
\end{aligned}$$

Since a_i is bounded by $A/h_N^{d_x}$, by choosing $h_N = \frac{1}{2}N^{-\frac{1}{2d_x+3}}$, the right-hand side is lower bounded by:

$$
\begin{aligned}
& N^{(2d_x+2)/(2d_x+3)} - \left|a_i - Nf_X(x) + (N-1)\mathbb{E}[a_j]\right| \\
\geq\ & N^{(2d_x+2)/(2d_x+3)} - |a_i - f_X(x)| - (N-1)|\mathbb{E}[a_j] - f_X(x)| \\
\geq\ & N^{(2d_x+2)/(2d_x+3)} - A/h_N^{d_x} - Nh_N \\
\geq\ & \frac{1}{3}N^{(2d_x+2)/(2d_x+3)}.
\end{aligned}
\tag{A.40}
$$

Since for $j \neq i$, $a_j's$ are i.i.d and bounded by $A/h_N^{d_x}$, by Hoeffding's inequality, we obtain

$$
\begin{aligned}
& \mathbb{P}\Big(\Big| \sum_{j\neq i} a_j - (N-1)\mathbb{E}[a_j] \Big| > N^{(2d_x+2)/(2d_x+3)} \\
& - \left|a_i - Nf_X(x) + (N-1)\mathbb{E}[a_j]\right| \Big| X_i = x \Big) \\
\leq\ & \mathbb{P}\Big(\Big| \sum_{j\neq i} a_j - (N-1)\mathbb{E}[a_j] \Big| > \frac{1}{3}N^{(2d_x+2)/(2d_x+3)} \Big| X_i = x \Big) \\
\leq\ & 2\exp\{-2\frac{(\frac{1}{3}N^{(2d_x+2)/(2d_x+3)})^2}{(N-1)\cdot A^2 h_N^{-2d_x}}\} \\
\leq\ & 2\exp\{-\frac{2}{9A^2(N-1)}(N^{(2d_x+2)/(2d_x+3)}h_N^{d_x})^2\} \\
\leq\ & 2\exp\{-\frac{N^{1/(2d_x+3)}}{9A^2}\}.
\end{aligned}
\tag{A.41}
$$

Since this upper bound is independent of x, we can take expectation over x to obtain the desired claim.

A.1.3 Proof of Lemma 2

Define

$$
\hat{f}_{X,Y}(X_i, Y_i) = \frac{\exp\{\psi(k) - \psi(N)\}}{c_{d_x+d_y}\rho_{k,i}^{d_x+d_y}}\ , \tag{A.42}
$$

so that

$$
\hat{H}_{k,N}^{U}(X,Y) = -\sum_{i=1}^{N} \frac{w_i'}{N} \log \hat{f}_{X,Y}(X_i, Y_i)\ . \tag{A.43}
$$

By Theorem 8 of [42], we have

$$\lim_{N\to\infty} \mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big] \;=\; \log f_{X,Y}(x,y)\;. \quad \text{(A.44)}$$

Notice that $w_i' \log \hat f(X_i,Y_i)$ are identically distributed, therefore, by plugging in $w_i' = u(X_i)/f_X(X_i)$, we have

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}\hat H^U_{k,N}(X,Y)\\
= &- \lim_{N\to\infty} \mathbb{E}[w_i' \log \hat f_{X,Y}(X_i,Y_i)]\\
= &- \lim_{N\to\infty} \iint \mathbb{E}\big[\frac{u(X_i)}{f_X(X_i)} \log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big] f_{X,Y}(x,y)dxdy\\
= &- \lim_{N\to\infty} \iint u(x) f_{Y|X}(y|x)\mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big]dxdy\;.
\end{aligned} \quad \text{(A.45)}$$

Now we want to show that

$$\begin{aligned}
&\lim_{N\to\infty} \iint u(x) f_{Y|X}(y|x)\mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big]dxdy\\
= &\iint u(x) f_{Y|X}(y|x)\big(\lim_{N\to\infty} \mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big]\big)dxdy\\
= &\iint u(x) f_{Y|X}(y|x) \log f_{X,Y}(x,y)dxdy\;,
\end{aligned} \quad \text{(A.46)}$$

which follows from the reverse Fatou's lemma and the fact that

$$\begin{aligned}
&\limsup_{N\to\infty} \iint \big|\frac{u(x)}{f(x)}\mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big]\big|^2 f(x,y)dxdy\\
\le\; &C_1^{-2} \limsup_{N\to\infty} \iint \big|\mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big]\big|^2 f(x,y)dxdy\\
\le\; &C_1^{-2} \iint \limsup_{N\to\infty} \big|\mathbb{E}\big[\log \hat f_{X,Y}(X_i,Y_i)\big|(X_i,Y_i)=(x,y)\big]\big|^2 f(x,y)dxdy\\
\le\; &C_1^{-2} \iint \limsup_{N\to\infty} \big(\log f_{X,Y}(x,y)\big)^2 f(x,y)dxdy\\
<\; &+\infty\;.
\end{aligned} \quad \text{(A.47)}$$

As explained in the main result section, we regularize the kNN distance such that $\rho_{k,i}^{d_x+d_y} > Ck/N$ for some positive constant C. This ensures that $\log \hat f_{X,Y}(X_i,Y_i) < C'$ almost surely. It follows that $\mathbb{E}[\log \hat f_{X,Y}(X_i,Y_i)|X_i = x, Y_i = y] < C'$ and one can apply reverse Fatou's lemma. Similar interchange of limit has been used in [10, 16] without the regularization; in this context [17] claims that this step is not justified (although no counterexample is pointed out). But in our case, given the practical way the algorithm is implemented with the regularization, reverse Fatou's lemma is justified. Therefore,

$$\lim_{N\to\infty} \mathbb{E}\hat{H}^U_{k,N}(X,Y) = -\iint u(x) f_{Y|X}(y|x) \log f_{X,Y}(x,y) dxdy. \tag{A.48}$$

Moreover, by Theorem 11 of [42], we have:

$$\lim_{N\to\infty} \mathrm{Var}[\hat{f}_{X,Y}(X_i,Y_i)] \;=\; (\frac{\Gamma'(k)}{\Gamma(k)})' \,\mathrm{Var}[\log f_{X,Y}(x,y)]\ , \tag{A.49}$$

and for any $j \neq i$:

$$\lim_{N\to\infty} \mathrm{Cov}[\hat{f}_{X,Y}(X_i,Y_i), \hat{f}_{X,Y}(X_j,Y_j)] = 0\ . \tag{A.50}$$

By $w'_i \leq 1/C_1$ for all i and the fact that $\hat{f}_{X,Y}(X_i,Y_i)$ are identically distributed, we have:

$$\begin{aligned}
&\mathrm{Var}\big[\hat{H}^U_{k,N}(X,Y)\big] \\
=\;& \sum_{i=1}^{N} \frac{(w'_i)^2}{N^2}\mathrm{Var}[\hat{f}_{X,Y}(X_i,Y_i)] + \sum_{j\neq i} \frac{w'_i w'_j}{N^2}\mathrm{Cov}[\hat{f}_{X,Y}(X_i,Y_i), \hat{f}_{X,Y}(X_j,Y_j)] \\
\leq\;& \sum_{i=1}^{N} \frac{1}{C_1^2 N^2}\mathrm{Var}[\hat{f}_{X,Y}(X_i,Y_i)] + \sum_{j\neq i} \frac{1}{C_1^2 N^2}\mathrm{Cov}[\hat{f}_{X,Y}(X_i,Y_i), \hat{f}_{X,Y}(X_j,Y_j)] \\
=\;& \frac{1}{C_1^2 N}((\frac{\Gamma'(k)}{\Gamma(k)})'\mathrm{Var}[\log f_{X,Y}(x,y)]) \\
&+ \frac{1}{C_1^2 N^2}\binom{N}{2}\mathrm{Cov}[\hat{f}_{X,Y}(X_1,Y_1), \hat{f}_{X,Y}(X_2,Y_2)]\ .
\end{aligned} \tag{A.51}$$

Therefore,

$$\lim_{N\to\infty} \mathrm{Var}\big[\hat{H}^U_{k,N}(X,Y)\big] = 0\ . \tag{A.52}$$

Combining (A.48) and (A.52), we get:

$$\begin{aligned}
& \lim_{N\to\infty} \mathbb{E}\Big[\big(\widehat{H}^U_{k,N}(X,Y) - \big(- \iint u(x) f_{Y|X}(y|x) \log f_{X,Y}(x,y) dxdy\,\big)\big)^2\Big] \\
= \; & \lim_{N\to\infty} \Big(\mathbb{E}\hat{H}^U_{k,N}(X,Y) - \big(- \iint u(x) f_{Y|X}(y|x) \log f_{X,Y}(x,y) dxdy\,\big)\Big)^2 \\
& + \lim_{N\to\infty} \mathrm{Var}\big[\hat{H}^U_{k,N}(X,Y)\big] \\
= \; & 0. \qquad \text{(A.53)}
\end{aligned}$$

Therefore, $\hat{H}^U_{k,N}(X,Y)$ converges to its mean in L^2, and hence in probability, i.e.,

$$\lim_{N\to\infty} \mathbb{P}\Big(\big|\widehat{H}^U_{k,N}(X,Y) + \iint u(x) f_{Y|X}(y|x) \log f_{X,Y}(x,y) dxdy\big| > \varepsilon \Big) = 0 \, . \qquad \text{(A.54)}$$

A.1.4 Proof of Lemma 3

Define

$$\hat{f}_X(X_i) \equiv \frac{n_{x,i}}{(N-1) c_{d_x} \rho^{d_x}_{k,i}} \, , \qquad \text{(A.55)}$$

$$\hat{f}_U(Y_i) \equiv \frac{n_{y,i}}{(N-1) c_{d_y} \rho^{d_y}_{k,i}} \, , \qquad \text{(A.56)}$$

such that

$$\hat{H}^U_{k,N}(X) + \hat{H}^U_{k,N}(Y) = -\sum_{i=1}^{N} \frac{w'_i}{N}\big(\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)\,\big). \qquad \text{(A.57)}$$

By triangle inequality, we can write the formula in Lemma 3 as:

$$
\begin{aligned}
& \Big| \iint u(x) f_{Y|X}(y|x)\big(\log f_X(x) + \log f_U(y)\big) dxdy + \widehat{H}^U_{k,N}(X) + \widehat{H}^U_{k,N}(Y) \Big| \\
= & \Big| \iint u(x) f_{Y|X}(y|x)\big(\log f_X(x) + \log f_U(y)\big) dxdy \\
& - \sum_{i=1}^{N} \frac{w'_i}{N}\big(\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i) \big) \Big| \\
\leq & \Big| \iint u(x) f_{Y|X}(y|x)\big(\log f_X(x) + \log f_U(y)\big) dxdy \\
& - \sum_{i=1}^{N} \frac{w'_i}{N}\big(\log f_X(X_i) + \log f_U(Y_i) \big) \Big| \qquad \text{(A.58)} \\
& + \sum_{i=1}^{N} \frac{w'_i}{N}\Big| \big(\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)\big) - \big(\log f_X(X_i) + \log f_U(Y_i)\big) \Big|. \qquad \text{(A.59)}
\end{aligned}
$$

The first term (A.58) comes from sampling. Recall that $w'_i = u(X_i)/f_X(X_i)$. Since the random variables $w'_i\big(\log f_X(X_i) + \log f_U(Y_i) \big)$ are i.i.d., therefore, by the strong law of large numbers,

$$
\sum_{i=1}^{N} \frac{w'_i}{N}\big(\log f_X(X_i) + \log f_U(Y_i) \big) \to \mathbb{E}\Big(\frac{u(x)}{f_X(x)}\big(\log f_X(x) + \log f_U(y) \big) \Big) \;, \tag{A.60}
$$

almost surely. The mean is given by

$$
\begin{aligned}
& \mathbb{E}\Big(\frac{u(x)}{f_X(x)}(\log f_X(x) + \log f_U(y)) \Big) \\
= & \iint \frac{u(x)}{f_X(x)}\big(\log f_X(x) + \log f_U(y) \big) f(x,y) dxdy \\
= & \iint u(x) f_{Y|X}(y|x)\big(\log f_X(x) + \log f_U(y) \big) dxdy. \qquad \text{(A.61)}
\end{aligned}
$$

Therefore, (A.58) converges to 0 almost surely.

The second term (A.59) comes from density estimation. To simplfy the notations, let $Z_i = (X_i, Y_i)$, $z = (x, y)$ and $f(z) = f(x, y)$. For any fixed

$\varepsilon > 0$, by union bound, we obtain that

$$
\begin{aligned}
& \mathbb{P}\big(\sum_{i=1}^{N} \frac{w_i'}{N} \big| (\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)) - (\log f_X(X_i) + \log f_U(Y_i)) \big| > \varepsilon \big) \\
\leq \; & \mathbb{P}\big(\bigcup_{i=1}^{N} \big\{ \big| (\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)) - (\log f_X(X_i) + \log f_U(Y_i)) \big| > \varepsilon/2 \big\} \big) \\
& + \mathbb{P}\big(\sum_{i=1}^{N} w_i' > 2N \big). \qquad \text{(A.62)}
\end{aligned}
$$

The second term converges to zero by Lemma 4. The first term is bounded by:

$$
\begin{aligned}
& \mathbb{P}\big(\bigcup_{i=1}^{N} \big\{ \big| (\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)) - (\log f_X(X_i) + \log f_U(Y_i)) \big| > \varepsilon/2 \big\} \big) \\
\leq \; & N \cdot \mathbb{P}\big(\big| (\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)) - (\log f_X(X_i) + \log f_U(Y_i)) \big| > \varepsilon/2 \big) \\
\leq \; & N \int \big(I_1(z) + I_2(z) + I_3(z) + I_4(z) \big) f(z) dz \qquad \text{(A.63)}
\end{aligned}
$$

where

$$
\begin{aligned}
I_1(z) &= \mathbb{P}\big(\rho_{k,i} > r_1 \big| Z_i = z \big) \, , && \text{(A.64)} \\
I_2(z) &= \mathbb{P}\big(\rho_{k,i} < r_2 \big| Z_i = z \big) \, , && \text{(A.65)} \\
I_3(z) &= \int_{r=r_2}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_X(X_i) - \log f_X(X_i) \big| > \varepsilon/4 \\
& \qquad \big| \, \rho_{k,i} = r, Z_i = z \big) f_{\rho_{k,i}}(r) dr \, , && \text{(A.66)} \\
I_4(z) &= \int_{r=r_2}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_U(Y_i) - \log f_U(Y_i) \big| > \varepsilon/4 \\
& \qquad \big| \, \rho_{k,i} = r, Z_i = z \big) f_{\rho_{k,i}}(r) dr \, , && \text{(A.67)}
\end{aligned}
$$

where $f_{\rho_{k,i}}(r)$ is the pdf of $\rho_{k,i}$ given $Z_i = z$. Here

$$
r_1 = \log N (N f(z) c_{d_x + d_y})^{-\frac{1}{d_x + d_y}} \qquad \text{(A.68)}
$$

$$
r_2 = \max \big\{ (\log N)^2 (N f_X(x) c_{d_x})^{-\frac{1}{d_x}}, (\log N)^2 (N f_U(y) c_{d_y})^{-\frac{1}{d_y}} \big\}. \qquad \text{(A.69)}
$$

$I_1(z)$ and $I_2(z)$ are the probability that the k-NN distance $\rho_{k,i}$ is large or small given $Z_i = z$. $I_3(z)$ and $I_4(z)$ gives the probability that the estimator deviates from the true value, given that $\rho_{k,i}$ is medium. We will consider the four terms separately.

I_1: Let $B_Z(z,r) = \{Z : \|Z-z\| < r\}$ be the (d_x+d_y)-dimensional ball centered at z with radius r. Since the Hessian matrix of $H(f)$ exists and $\|H(f)\|_2 < C$ almost everywhere, then for sufficiently small r, the probability mass within $B_Z(z,r)$ is given by

$$
\begin{aligned}
&\mathbb{P}\big(\, u \in B_Z(z,r)\,\big) = \int_{\|u-z\|\le r} f(u)du \\
= \quad & \int_{\|u-z\|\le r} f(z) + (u-z)^T \nabla f(z) + (u-z)^T H(f)(z)(u-z)du \\
\in \quad & \big[\, f(z) c_{d_x+d_y} r^{d_x+d_y}(1-Cr^2)), f(z) c_{d_x+d_y} r^{d_x+d_y}(1+Cr^2))\,\big]. \qquad \text{(A.70)}
\end{aligned}
$$

Then for sufficiently large N, the probability mass within $B_Z(z,r_1)$ is lower bounded by

$$
\begin{aligned}
p_1 &\equiv \mathbb{P}\big(\, u \in B_Z(z,r_1)\,\big) \\
&\ge f(z) c_{d_x+d_y} r_1^{d_x+d_y}\big(\, 1 - Cr_1^2\,\big) \\
&\ge \frac{(\log N)^{d_x+d_y}}{2N}. \qquad \text{(A.71)}
\end{aligned}
$$

$I_1(z)$ is the probability that at most k samples fall in $B_Z(z,r_1)$, so it is upper bounded by

$$
\begin{aligned}
I_1(z) &= \mathbb{P}\big(\, \rho_{k,i} > r_1 \big| Z_i = z\,\big) \\
&= \sum_{m=0}^{k-1} \binom{N-1}{m} p_1^m (1-p_1)^{N-1-m} \\
&\le \sum_{m=0}^{k-1} N^m (1-p_1)^{N-1-m} \\
&\le kN^{k-1}(1 - \frac{(\log N)^{d_x+d_y}}{2N})^{N-k-1} \\
&\le kN^{k-1}\exp\{-\frac{(\log N)^{d_x+d_y}(N-k-1)}{2N}\} \\
&\le kN^{k-1}\exp\{-\frac{(\log N)^{d_x+d_y}}{4}\}\,, \qquad \text{(A.72)}
\end{aligned}
$$

for any $d_x, d_y \ge 1$.

I_2**:** Let $r_{2,1} \equiv (\log N)^2 (N f_X(x) c_{d_x})^{-\frac{1}{d_x}}$. Then for sufficiently large N, the probability mass within $B_Z(z, r_{2,1})$ is given by:

$$\begin{aligned}
p_{2,1} &\equiv \mathbb{P}\big(\, u \in B_Z(z, r_{2,1})\,\big) \\
&\le f(z) c_{d_x+d_y} r_{2,1}^{d_x+d_y} \big(\, 1 + C r_{2,1}^2 \,\big) \\
&\le \frac{2 f(z) c_{d_x+d_y}}{\big(f(x) c_{d_x}\big)^{\frac{d_x+d_y}{d_x}}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_x}} \\
&\le 2 f_{Y|X}(y|x) \frac{c_{d_x+d_y}}{c_{d_x}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_x}} \\
&\le 2C' \frac{c_{d_x+d_y}}{c_{d_x}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_x}} \ ,
\end{aligned} \tag{A.73}$$

where the last equation comes from the assumption that $f_{Y|X}(y|x) < C'$. Similarly, let $r_{2,2} = (\log N)^2 (N f_U(y) c_{d_y})^{-\frac{1}{d_y}}$, the probability of being in $B_Z(z, r_{2,2})$ is

$$\begin{aligned}
p_{2,2} &\equiv \mathbb{P}\big(\, u \in B_Z(z, r_{2,2})\,\big) \\
&\le f(z) c_{d_x+d_y} r_{2,2}^{d_x+d_y} \big(\, 1 + C r_{2,2}^2 \,\big) \\
&\le \frac{2 f(z) c_{d_x+d_y}}{\big(f_U(y) c_{d_y}\big)^{\frac{d_x+d_y}{d_y}}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_y}} \\
&\le 2 \frac{f(z)}{f_U(y)} \frac{c_{d_x+d_y}}{c_{d_y}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_y}} \\
&\le 2C_2 \frac{f_U(x,y)}{f_U(y)} \frac{c_{d_x+d_y}}{c_{d_y}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_y}} \\
&\le 2C_2 C' \frac{c_{d_x+d_y}}{c_{d_y}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_y}} .
\end{aligned} \tag{A.74}$$

$I_2(z)$ is the probability that at least k samples lie in $B_Z(z, \max\{r_{2,1}, r_{2,2}\})$.

It is upper bounded as follows:

$$
\begin{aligned}
I_2(z) &= \mathbb{P}\big(\rho_{k,i} < \max\big\{ r_{2,1}, r_{2,2} \big\} \big| Z_i = z \big) \\
&= \sum_{m=k}^{N-1} \binom{N-1}{m} \max\{p_{2,1}, p_{2,2}\}^m (1 - \{p_{2,1}, p_{2,2}\})^{N-1-m} \\
&\leq \sum_{m=k}^{N-1} N^m \max\{p_{2,1}, p_{2,2}\}^m \\
&\leq \sum_{m=k}^{N-1} \big(2C'C_2 \frac{c_{d_x+d_y}}{\min\{c_{d_x}, c_{d_y}\}} (\log N)^{2(d_x+d_y)} N^{-\min\{\frac{d_y}{d_x}, \frac{d_x}{d_y}\}} \big)^m \\
&\leq (4C'C_2 \frac{c_{d_x+d_y}}{\min\{c_{d_x}, c_{d_y}\}})^k (\log N)^{2k(d_x+d_y)} N^{-k\min\{\frac{d_y}{d_x}, \frac{d_x}{d_y}\}} \;, \quad \text{(A.75)}
\end{aligned}
$$

for sufficiently large N such that $\frac{c_{d_x+d_y}}{\min\{c_{d_x}, c_{d_y}\}} (\log N)^{2(d_x+d_y)} N^{-\min\{\frac{d_x+d_y}{d_x}, \frac{d_x+d_y}{d_y}\}} < 1/(4C')$, the last inequality comes from sum of geometric series. This holds for any $d_x, d_y \geq 1$ and $k \geq 1$.

I_3**:** Given that $Z_i = z = (x, y)$ and $\rho_{k,i} = r$. Recall that $\hat{f}_X(X_i) = \frac{n_{x,i}}{(N-1)c_{d_x}\rho_{k,i}^{d_x}}$, so we have

$$
\begin{aligned}
& \mathbb{P}\big(\big| \log \hat{f}_X(X_i) - \log f_X(X_i) \big| > \varepsilon/4 \big| \rho_{k,i} = r, Z_i = z \big) \\
= \; & \mathbb{P}\big(\big| \log n_{x,i} - \log (N-1) c_{d_x} r^{d_x} f_X(x) \big| > \varepsilon/4 \big| \rho_{k,i} = r, Z_i = z \big) \\
= \; & \mathbb{P}\big(n_{x,i} > (N-1) c_{d_x} r^{d_x} f_X(x) e^{\varepsilon/4} \big| \rho_{k,i} = r, Z_i = z \big) \\
& + \mathbb{P}\big(n_{x,i} < (N-1) c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} \big| \rho_{k,i} = r, Z_i = z \big). \qquad \text{(A.76)}
\end{aligned}
$$

Given $Z_i = z$ and $\rho_{k,i} = r \in [r_{2,1}, r_1]$, the probability distribution of $n_{x,i}$ is given in the following lemma.

Lemma 6. *Given $Z_i = z = (x, y)$ and $\rho_{k,i} = r < r_N$ for some deterministic sequence of r_N such that $\lim_{N\to\infty} r_N = 0$ and for any positive $\varepsilon > 0$, the number of neighbors $n_{x,i} - k$ is distributed as $\sum_{l=k+1}^{N-1} U_l$, where U_l are i.i.d Bernoulli random variables with mean $f_X(x) c_{d_x} r^{d_x} (1 - \varepsilon/8) \leq \mathbb{E}[U_l] \leq f_X(x) c_{d_x} r^{d_x} (1 + \varepsilon/8)$ for sufficiently large N.*

Given Lemma. 6, we obtain

$$\begin{aligned}
&\mathbb{P}\Big(n_{x,i} > (N-1)c_{d_x} r^{d_x} f_X(x) e^{\varepsilon/4} \Big| \rho_{k,i} = r, Z_i = z \Big) \\
= \; & \mathbb{P}\Big(\sum_{l=k+1}^{N-1} U_l > (N-1)c_{d_x} r^{d_x} f_X(x) e^{\varepsilon/4} - k \Big) \\
= \; & \mathbb{P}\Big(\sum_{l=k+1}^{N-1} U_l - N'\mathbb{E}[U_l] > (N-1)c_{d_x} r^{d_x} f_X(x) e^{\varepsilon/4} - k - N'\mathbb{E}[U_l] \Big) \,,
\end{aligned} \tag{A.77}$$

where $N' = N - k - 1$ for simplicity. Here the right-hand side in the probability is lower bounded by

$$\begin{aligned}
& (N-1)c_{d_x} r^{d_x} f_X(x) e^{\varepsilon/4} - k - \mathbb{E}[U_l] \\
\geq \; & (N-1)c_{d_x} r^{d_x} f_X(x) e^{\varepsilon/4} - k - N' f_X(x) c_{d_x} r^{d_x} (1 + \varepsilon/8) \\
\geq \; & N' c_{d_x} r^{d_x} f_X(x) (e^{\varepsilon/4} - 1 - \varepsilon/8) - k \\
\geq \; & N' c_{d_x} r^{d_x} f_X(x) \varepsilon/16 \,,
\end{aligned} \tag{A.78}$$

for sufficiently large N such that $N' c_{d_x} r^{d_x} f_X(x)(e^{\varepsilon/4} - 1 - \varepsilon/16) > k$. Since U_l is Bernoulli, we have $\mathbb{E}[U_l^2] = \mathbb{E}[U_l]$. Now applying Bernstein's inequality, (A.77) is upper bounded by:

$$\begin{aligned}
& \mathbb{P}\Big(\sum_{l=k+1}^{N-1} U_l - N'\mathbb{E}[U_l] > (N-1)c_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - k - N'\mathbb{E}[U_l] \Big) \\
\leq \; & \mathbb{P}\Big(\sum_{l=k+1}^{N-1} U_l - N'\mathbb{E}[U_l] > N' c_{d_x} r^{d_x} f_X(x) \varepsilon/16 \Big) \\
\leq \; & \exp\{ - \frac{(N' c_{d_x} r^{d_x} f_X(x) \varepsilon/16)^2}{2\big(N'\mathbb{E}[U_l^2] + \frac{1}{3}(N' c_{d_x} r^{d_x} f_X(x)\varepsilon/16) \big)} \} \\
\leq \; & \exp\{ - \frac{(N' c_{d_x} r^{d_x} f_X(x) \varepsilon/16)^2}{2\big(N' c_{d_x} r^{d_x} f_X(x)(1+\varepsilon/8) + \frac{1}{3}(N' c_{d_x} r^{d_x} f_X(x)\varepsilon/16) \big)} \} \\
= \; & \exp\{ - \frac{\varepsilon^2}{512(1 + 7\varepsilon/48)} N' c_{d_x} r^{d_x} f_X(x) \} \,.
\end{aligned} \tag{A.79}$$

Similarly, the tail bound on the other direction is given by:

$$
\begin{aligned}
& \mathbb{P}\big(\, n_{x,i} < (N-1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} \Big| \rho_{k,i} = r, Z_i = z \,\big) \\
= \; & \mathbb{P}\big(\sum_{l=k+1}^{N-1} U_l < (N-1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} - k \big) \\
= \; & \mathbb{P}\big(\sum_{l=k+1}^{N-1} U_l - N'\mathbb{E}[U_l] < (N-1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} - k - N'\mathbb{E}[U_l] \big) \;,
\end{aligned} \tag{A.80}
$$

where the right-hand side is negative and upper bounded by:

$$
\begin{aligned}
& (N-1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} - k - \mathbb{E}[U_l] \\
\leq \; & (N-1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} - k - (N-k-1) f_X(x) c_{d_x} r^{d_x} (1-\varepsilon/8) \\
\leq \; & N' c_{d_x} r^{d_x} f_X(x) (e^{-\varepsilon/4} - 1 + \varepsilon/8) \\
\leq \; & -N' c_{d_x} r^{d_x} f_X(x) \varepsilon/16 \;,
\end{aligned} \tag{A.81}
$$

for small enough r such that $c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} \leq 1$ and small enough ε that $e^{-\varepsilon/4} - 1 + 3\varepsilon/16 < 0$. Similarly, (A.80) is upper bounded by:

$$
\begin{aligned}
& \mathbb{P}\big(\sum_{l=k+1}^{N-1} U_l - N'\mathbb{E}[U_l] < (N-1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon/4} - k - N'\mathbb{E}[U_l] \big) \\
\leq \; & \exp\{ -\frac{\varepsilon^2}{512(1+7\varepsilon/48)} N' c_{d_x} r^{d_x} f_X(x) \}.
\end{aligned} \tag{A.82}
$$

Therefore, $I_3(z)$ is upper bounded by:

$$
\begin{aligned}
I_3(z) \; = \; & \int_{r=r_2}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_X(X_i) - \log f_X(X_i) \big| > \varepsilon \big| \rho_{k,i} = r, Z_i = z \,\big) f_{\rho_{k,i}}(r) dr \\
\leq \; & \int_{r=r_{2,1}}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_X(X_i) - \log f_X(X_i) \big| > \varepsilon \big| \rho_{k,i} = r, Z_i = z \,\big) f_{\rho_{k,i}}(r) dr \\
\leq \; & \int_{r=r_{2,1}}^{r_1} 2\exp\{ -\frac{\varepsilon^2}{512(1+7\varepsilon/48)} N' c_{d_x} r^{d_x} f_X(x) \} f_{\rho_{k,i}}(r) dr \\
\leq \; & 2\exp\{ -\frac{\varepsilon^2}{1024} N c_{d_x} f_X(x) ((\log N)^2 (N f_X(x) c_{d_x})^{-\frac{1}{d_x}})^{d_x} \} \\
\leq \; & 2\exp\{ -\frac{\varepsilon^2}{1024} (\log N)^{2d_x} \} \;,
\end{aligned} \tag{A.83}
$$

for sufficiently large N such that $N'/(1+\frac{7}{48}\varepsilon) > N/2$.

I_4**:** Given that $Z_i = z = (x,y)$ and $\rho_{k,i} = r$. Recall that $\hat{f}_U(Y_i) = \frac{n_{y,i}}{(N-1)c_{d_y} r^{d_y}}$, then we have

$$\begin{aligned}
& \mathbb{P}\Big(\big| \log \hat{f}_U(Y_i) - \log f_U(Y_i) \big| > \varepsilon/4 \big| \rho_{k,i} = r, Z_i = z \Big) \\
= \; & \mathbb{P}\Big(\big| \log n_{y,i} - \log (N-1) c_{d_y} r^{d_y} f_Y(y) \big| > \varepsilon/4 \big| \rho_{k,i} = r, Z_i = z \Big) \\
= \; & \mathbb{P}\Big(n_{y,i} > (N-1) c_{d_y} r^{d_y} f_U(y) e^{\varepsilon/4} \big| \rho_{k,i} = r, Z_i = z \Big) \\
& + \mathbb{P}\Big(n_{y,i} < (N-1) c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} \big| \rho_{k,i} = r, Z_i = z \Big). \qquad \text{(A.84)}
\end{aligned}$$

Recall that

$$n_{y,i} = \sum_{j \neq i} w'_j \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\} = \sum_{j \neq i} \frac{u(X_i)}{f_X(X_i)} \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \text{ (A.85)}$$

We write $n_{y,i} = m_{y,i}^{(1)} + m_{y,i}^{(2)}$, where

$$\begin{aligned}
m_{y,i}^{(1)} &= \sum_{j : \|Z_j - z\| < \rho_{k,i}} \frac{u(X_j)}{f_X(X_j)} \\
m_{y,i}^{(2)} &= \sum_{j : \|Z_j - z\| > \rho_{k,i}} \frac{u(X_j)}{f_X(X_j)} \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \qquad \text{(A.86)}
\end{aligned}$$

Since $C_1/\mu(K) \leq f_X(X_j) \leq C_2/\mu(K)$, we have: $k/C_2 \leq m_{y,i}^{(1)} \leq k/C_1$. Given that $Z_i = z$ and $\rho_{k,i} = r \in [r_{2,2}, r_1]$, the probability distribution of $m_{y,i}^{(2)}$ is given by the following lemma.

Lemma 7. *Given $Z_i = z = (x,y)$ and $\rho_{k,i} = r < r_N$ for some deterministic sequence of r_N such that $\lim_{N\to\infty} r_N = 0$ and for a positive $\varepsilon > 0$, the distribution of $m_{y,i}^{(2)}$ is distributed as $\sum_{l=k+1}^{N-1} V_l$ where V_l are i.i.d random variables with $V_l \in [0, 1/C_1]$ and mean $f_U(y) c_{d_y} r^{d_y} (1-\varepsilon/8) \leq \mathbb{E}[V_l] \leq f_U(y) c_{d_y} r^{d_y} (1+\varepsilon/8)$, for sufficiently large N.*

Given Lemma 7 and the fact that $m_{y,i}^{(1)} \geq k/C_2$, we obtain

$$
\begin{aligned}
& \mathbb{P}\big(n_{y,i} > (N-1)c_{d_y} r^{d_y} f_U(y) e^{\varepsilon/4} \Big| \rho_{k,i} = r, Z_i = z \big) \\
\leq \;& \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l > (N-1)c_{d_y} r^{d_y} f_U(y) e^{\varepsilon/4} - k/C_2 \big) \\
= \;& \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l - N'\mathbb{E}[V_l] > (N-1)c_{d_y} r^{d_y} f_U(y) e^{\varepsilon/4} - k/C_2 - N'\mathbb{E}[V_l] \big) \ ,
\end{aligned}
\tag{A.87}
$$

here the right-hand side is lower bounded by

$$
\begin{aligned}
& (N-1)c_{d_y} r^{d_y} f_U(y) e^{\varepsilon/4} - k/C_2 - \mathbb{E}[V_l] \\
\geq \;& (N-1)c_{d_y} r^{d_y} f_U(y) e^{\varepsilon/4} - k/C_2 - N'c_{d_y} r^{d_y} f_U(y)(1+\varepsilon/8) \\
\geq \;& N'c_{d_y} r^{d_y} f_U(y)(e^{\varepsilon/4} - 1 - \varepsilon/8) - k/C_2 \\
\geq \;& N'c_{d_y} r^{d_y} f_U(y)\varepsilon/16 \ ,
\end{aligned}
\tag{A.88}
$$

for sufficiently large N such that $N'c_{d_y} r^{d_y} f_U(y)(e^{\varepsilon/4} - 1 - \varepsilon/16) > k/C_2$. Since V_l is upper bounded by $1/C_1$, so $\mathbb{E}[V_l^2] \leq \mathbb{E}[V_l]/C_1$. Now applying Bernstein's inequality, (A.87) is upper bounded by:

$$
\begin{aligned}
& Pr\big(\sum_{l=k+1}^{N-1} V_l - N'\mathbb{E}[V_l] > (N-1)c_{d_y} r^{d_y} f_U(y) e^{\varepsilon} - k - N'\mathbb{E}[V_l] \big) \\
\leq \;& \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l - N'\mathbb{E}[V_l] > N'c_{d_y} r^{d_y} f_U(y)\varepsilon/16 \big) \\
\leq \;& \exp\{ - \frac{(N'c_{d_y} r^{d_y} f_U(y)\varepsilon/16)^2}{2\big(N'\mathbb{E}[V_l^2] + \frac{1}{3C_1}(N'c_{d_y} r^{d_y} f_U(y)\varepsilon/8) \big)} \} \\
\leq \;& \exp\{ - \frac{(N'c_{d_y} r^{d_y} f_U(y)\varepsilon/16)^2}{2\big(N'c_{d_y} r^{d_y} f_U(y)(1+\varepsilon/8)/C_1 + \frac{1}{3C_1}(N'c_{d_y} r^{d_y} f_U(y)\varepsilon/16) \big)} \} \\
\leq \;& \exp\{ - \frac{C_1 \varepsilon^2}{512(1+7\varepsilon/48)} N'c_{d_y} r^{d_y} f_U(y) \} \ .
\end{aligned}
\tag{A.89}
$$

Similarly, since $m_{y,i}^{(1)} < k/C_1$, the tail bound on the other way is given by:

$$\begin{aligned}
& \mathbb{P}\big(n_{y,i} < (N-1)c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} \big| \rho_{k,i} = r, Z_i = z \big) \\
\leq\ & \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l < (N-1)c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} - k/C_1 \big) \\
=\ & \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l - N'\mathbb{E}[V_l] < (N-1)c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} - k/C_1 - N'\mathbb{E}[V_l] \big) ,
\end{aligned} \tag{A.90}$$

where the right-hand side is negative and upper bounded by:

$$\begin{aligned}
& (N-1)c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} - k/C_1 - \mathbb{E}[V_l] \\
\leq\ & (N-1)c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} - k/C_1 - (N-k-1)c_{d_y} r^{d_y} f_U(y)(1-\varepsilon/8) \\
\leq\ & N' c_{d_y} r^{d_y} f_U(y)(e^{-\varepsilon/4} - 1 + \varepsilon/8) \\
\leq\ & -N' c_{d_y} r^{d_y} f_U(y)\varepsilon/16 ,
\end{aligned} \tag{A.91}$$

for small enough r such that $c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} \leq 1/C_1$ and small enough ε that $e^{-\varepsilon/4} - 1 + 3\varepsilon/16 < 0$. Similarly, (A.90) is upper bounded by:

$$\begin{aligned}
& \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l - N'\mathbb{E}[V_l] < (N-1)c_{d_y} r^{d_y} f_U(y) e^{-\varepsilon/4} - k - N'\mathbb{E}[V_l] \big) \\
\leq\ & \exp\{-\frac{C_1\varepsilon^2}{512(1+7\varepsilon/48)} N' c_{d_y} r^{d_y} f_U(y)\}.
\end{aligned} \tag{A.92}$$

Therefore, $I_4(z)$ is upper bounded by:

$$\begin{aligned}
I_4(z) &= \int_{r=r_2}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_U(Y_i) - \log f_U(Y_i) \big| > \varepsilon \big| \rho_{k,i} = r, Z_i = z \big) f_{\rho_{k,i}}(r) dr \\
&\leq \int_{r=r_{2,2}}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_U(Y_i) - \log f_U(Y_i) \big| > \varepsilon \big| \rho_{k,i} = r, Z_i = z \big) f_{\rho_{k,i}}(r) dr \\
&\leq \int_{r=r_{2,2}}^{r_1} \exp\{-\frac{C_1\varepsilon^2}{512(1+7\varepsilon/48)} N' c_{d_y} r^{d_y} f_U(y)\} f_{\rho_{k,i}}(r) dr \\
&\leq 2\exp\{-\frac{C_1\varepsilon^2}{1024} N c_{d_y} f_U(y) ((\log N)^2 (N f_U(y) c_{d_y})^{-\frac{1}{d_y}})^{d_y}\} \\
&\leq 2\exp\{-\frac{C_1\varepsilon^2}{1024} (\log N)^{2d_y}\} ,
\end{aligned} \tag{A.93}$$

for sufficiently large N such that $N'/(1+7\varepsilon/48)) > N/2$.

Now combining (A.72), (A.75), (A.83) and (A.93), we obtain

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^{N} \frac{w_i'}{N} \big|(\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)) - (\log f_X(X_i) + \log f_U(Y_i))\big| > \varepsilon \Big) \\
\leq\ & N \iint (I_1(z) + I_2(z) + I_3(z) + I_4(z)) f(z) dz \\
\leq\ & kN^k \exp\{-\frac{(\log N)^{d_x+d_y}}{4}\} \\
&+ (4C'C_2 \frac{c_{d_x+d_y}}{\min\{c_{d_x}, c_{d_y}\}})^k (\log N)^{2k(d_x+d_y)} N^{1-k\min\{\frac{d_y}{d_x}, \frac{d_y}{d_y}\}} \\
&+ 2N \exp\{-\frac{\varepsilon^2}{1024}(\log N)^{2d_x}\} + 2N \exp\{-\frac{C_1^2\varepsilon^2}{1024}(\log N)^{2d_y}\}. \qquad \text{(A.94)}
\end{aligned}$$

If $k > \max\{d_y/d_x, d_x/d_y\}$, we have $1 - k\min\{\frac{d_x+d_y}{d_x}, \frac{d_x+d_y}{d_y}\} < 0$. Then each of the four terms goes to 0 as $N \to \infty$ and we conclude:

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{P}\Big(\sum_{i=1}^{N} \frac{w_i'}{N} \big|&(\log \hat{f}_X(X_i) + \log \hat{f}_U(Y_i)) \\
&-(\log f_X(X_i) + \log f_U(Y_i))\big| > \varepsilon \Big) = 0\,. \qquad \text{(A.95)}
\end{aligned}$$

Therefore, by combining the convergence properties of error from kernel density estimation, error from self-normalized importance sampling and error from density estimation, we obtain that $\hat{I}^U_{k,N}(X,Y)$ converges to $I^U(f_{Y|X})$ in probability.

A.1.5 Proof of Lemma 6

Given that $Z_i = z = (x,y)$ and $\rho_{k,i} = r$, let $\{1, 2, \dots, i-1, i+1, \dots, N\} = S \cup \{j\} \cup T$ be a partition of the indexes with $|S| = k-1$ and $|T| = N-k-1$. Then define an event $\mathcal{A}_{S,j,T}$ associated to the partition as:

$$\mathcal{A}_{S,j,T} = \big\{\, \|Z_s - z\| < \|Z_j - z\|, \forall s \in S, \|Z_t - z\| > \|Z_j - z\|, \forall t \in T \,\big\}\,. \qquad \text{(A.96)}$$

Since $Z_j - z$ are i.i.d. random variables, each event $\mathcal{A}_{S,j,T}$ has identical probability. The number of such partitions is $\frac{(N-1)!}{(N-k-1)!(k-1)!}$, and thus $\mathbb{P}\big(\mathcal{A}_{S,j,T}\big) =$

$\frac{(N-k-1)!(k-1)!}{(N-1)!}$. So the cdf of $n_{x,i}$ is given by:

$$\begin{aligned}
& \mathbb{P}\big(\, n_{x,i} \leq k+m \big| \rho_{k,i} = r, Z_i = z \,\big) \\
= \ & \sum_{S,j,T} \mathbb{P}\big(\, \mathcal{A}_{S,j,T} \,\big) \mathbb{P}\big(\, n_{x,i} \leq k+m \big| \mathcal{A}_{S,j,T}, \rho_{k,i} = r, Z_i = z \,\big) \\
= \ & \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \mathbb{P}\big(\, n_{x,i} \leq k+m \big| \mathcal{A}_{S,j,T}, \rho_{k,i} = r, Z_i = z \,\big) \, .
\end{aligned} \tag{A.97}$$

Now condition on event $\mathcal{A}_{S,j,T}$ and $\rho_{k,i} = r$, namely Z_j is the k-nearest-neighbor with distance r, S is the set of samples with distance smaller than r and T is the set of samples with distance greater than r. Recall that $n_{x,i}$ is the number of samples with $\|X_j - x\| < r$. For any index $s \in S \cup \{j\}$, $\|X_s - x\| < r$ is satisfied. Therefore, $n_{x,i} \leq k+m$ means that there are no more than m samples in T with X-distance smaller than r. Let $U_l = \mathbb{I}\{\|X_l - x\| < r \big| \|Z_l - z\| > r\}$. Therefore,

$$\begin{aligned}
& \mathbb{P}\big(\, n_{x,i} \leq k+m \big| \mathcal{A}_{S,j,T}, \rho_{k,i} = r, Z_i = z \,\big) \\
= \ & \mathbb{P}\big(\, \sum_{t \in T} \mathbb{I}\{\|X_t - x\| < r\} \leq m \big| \ \|Z_s - z\| < r, \forall s \in S, \\
& \|Z_j - z\| = r, \|Z_t - z\| > r, \forall t \in T, Z_i = z \,\big) \\
= \ & \mathbb{P}\big(\, \sum_{t \in T} \mathbb{I}\{\|X_t - x\| < r\} \leq m \big| \ \|Z_t - z\| > r, \forall t \in T \,\big) \\
= \ & \mathbb{P}\big(\, \sum_{l=k+1}^{N-1} U_l \leq m \,\big).
\end{aligned} \tag{A.98}$$

We can drop the conditions of Z_s's for $s \notin T$ since Z_s and X_t are independent. Therefore, given that $\|Z_t - z\| > r$ for all $t \in T$, the variables $\mathbb{I}\{\|X_t - x\| < r\}$ are i.i.d. and have the same distribution as U_l. Therefore, we have:

$$
\begin{aligned}
& \mathbb{P}\big(\, n_{x,i} \le k+m \big| \rho_{k,i}=r, Z_i=z \,\big) \\
= \;& \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \mathbb{P}\big(\, n_{x,i} \le k+m \big| \mathcal{A}_{S,j,T}, \rho_{k,i}=r, Z_i=z \,\big) \\
= \;& \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \mathbb{P}\big(\, \sum_{l=k+1}^{N} U_l \le m \,\big) \\
= \;& \mathbb{P}\big(\, \sum_{l=k+1}^{N-1} U_l \le m \,\big) \;, \qquad \text{(A.99)}
\end{aligned}
$$

and so $n_{x,i}-k$ have the same distribution as $\sum_{l=k+1}^{N-1} U_l$ given $Z_i = z$ and $\rho_{k,i} = r$. Here the mean of U_l is given by:

$$
\begin{aligned}
\mathbb{E}[U_l] \;&=\; \mathbb{P}\big(\, \|X_l - x\| < r \big|\, \|Z_l - z\| > r \,\big) = \frac{\mathbb{P}\big(\, \|X_l - x\| < r, \|Z_l - z\| > r \,\big)}{\mathbb{P}\big(\, \|Z_l - z\| > r \,\big)} \\
&=\; \frac{\int_{\|u-x\|<r} f_X(u)du - \iint_{\|(u,v)-(x,y)\|,r} f(u,v)dudv}{1 - \iint_{\|(u,v)-(x,y)\|,r} f(u,v)dudv} . \qquad \text{(A.100)}
\end{aligned}
$$

Since $\|H(f_X)\| \le C$ almost everywhere, if $r < r_N$ and r_N decays as N goes to infinity, for sufficiently large N, we have the following bound for $\mathbb{E}[U_l]$:

$$
\begin{aligned}
\mathbb{E}[U_l] \;&<\; \int_{\|u-x\|<r} f_X(u)du \\
&=\; \int_{\|u-x\|<r} f_X(x) + (u-x)\nabla f_X(x) + (u-x)^T H(f_X)(x)(u-x) \\
&<\; f_X(x) c_{d_x} r^{d_x} (1 + Cr^2) \\
&<\; f_X(x) c_{d_x} r^{d_x} (1 + \varepsilon/8) \;, \qquad \text{(A.101)}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[U_l] \;&>\; \int_{\|u-x\|<r} f_X(u)du - \int_{\|(u,v)-(x,y)\|,r} f(u,v)dudv \\
&>\; f_X(x) c_{d_x} r^{d_x} (1 - Cr^2) - f(x,y) c_{d_x+d_y} r^{d_x+d_y} (1 + Cr^2) \\
&>\; f_X(x) c_{d_x} r^{d_x} (1 - \varepsilon/8). \qquad \text{(A.102)}
\end{aligned}
$$

A.1.6 Proof of Lemma 7

Given that $Z_i = z = (x, y)$ and $\rho_{k,i} = r$. Define $\mathcal{A}_{S,j,T}$ as same as in Lemma 6. Let $V_l = \frac{u(X_l)}{f_X(X_l)} \mathbb{I}\{\|Y_l - y\| < r \big| \|Z_l - z\| > r\}$. Condition on event $\mathcal{A}_{S,j,T}$, the CDF of $m^{(2)}_{y,i}$ is given by:,

$$
\begin{aligned}
& \mathbb{P}\big(m^{(2)}_{y,i} \le m \big| \mathcal{A}_{S,j,T}, \rho_{k,i} = r, Z_i = z \big) \\
= \ & \mathbb{P}\big(\sum_{t \in T} \frac{u(X_t)}{f_X(X_t)} \mathbb{I}\{\|Y_t - y\| < \rho_{k,i}\} \le m \big| \ \|Z_s - z\| < r, \forall s \in S, \\
& \|Z_j - z\| = r, \|Z_t - z\| > r, \forall t \in T, Z_i = z \big) \\
= \ & \mathbb{P}\big(\sum_{t \in T} \frac{u(X_t)}{f_X(X_t)} \mathbb{I}\{\|Y_t - y\| < \rho_{k,i}\} \le m \big| \ \|Z_t - z\| > r, \forall t \in T \big) \\
= \ & \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l \le m \big).
\end{aligned} \tag{A.103}
$$

Similarly we can drop the conditions of Z_s's for $s \notin T$. Therefore, given that $\|Z_t - z\| > r$ for all $t \in T$, the variables $\frac{u(X_t)}{f_X(X_t)} \mathbb{I}\{\|Y_t - y\| < r\}$ are i.i.d. and have the same distribution as V_l. Therefore, we have:

$$
\begin{aligned}
& \mathbb{P}\big(m^{(2)}_{y,i} \le m \big| \rho_{k,i} = r, Z_i = z \big) \\
= \ & \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \mathbb{P}\big(m^{(2)}_{y,i} \le m \big| \mathcal{A}_{S,j,T}, \rho_{k,i} = r, Z_i = z \big) \\
= \ & \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l \le m \big) \\
= \ & \mathbb{P}\big(\sum_{l=k+1}^{N-1} V_l \le m \big) ,
\end{aligned} \tag{A.104}
$$

so $m^{(2)}_{y,i}$ have the same distribution as $\sum_{l=k+1}^{N-1} V_l$ given $Z_i = z$ and $\rho_{k,i} = r$. Here $V_l \le \sup_x \frac{u(x)}{f_X(x)} = 1/C_1$. The mean of V_l is given by:

$$
\begin{aligned}
\mathbb{E}[V_l] &= \mathbb{E}\big[\frac{u(X_l)}{f_X(X_l)} \mathbb{I}\{\|Y_l - y\| < r\} \big| \ \|Z_l - z\| > r \big] \\
&= \frac{\iint_{\|v-y\|<r} \frac{u(u)}{f_X(u)} f(u,v) du dv - \iint_{\|(u,v)-(x,y)\|<r} \frac{u(u)}{f_X(u)} f(u,v) du dv}{1 - \iint_{\|(u,v)-(x,y)\|,r} \frac{u(u)}{f_X(u)} f(u,v) du dv} .
\end{aligned} \tag{A.105}
$$

Since $\|H(f_U)\| \leq C$ almost everywhere, if $r < r_N$ and r_N decays as N goes to infinity, for sufficiently large N, we have the following bound for $\mathbb{E}[V_l]$:

$$
\begin{aligned}
\mathbb{E}[V_l] &< \iint_{\|v-y\|<r} \frac{u(u)}{f_X(u)} f(u,v)dudv \\
&= \int_{\|v-y\|<r} f_U(v)dy \\
&= \int_{\|v-y\|<r} f_U(y) + (v-y)\nabla f_U(y) + (v-y)^T H(f_U)(y)(v-y) \\
&< f_U(y)c_{d_y} r^{d_y}(1 + Cr^2) \\
&< f_U(y)c_{d_y} r^{d_y}(1 + \varepsilon/8) \ , \qquad \text{(A.106)}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[V_l] &> \iint_{\|v-y\|<r} \frac{u(u)}{f_X(u)} f(u,v)dudv - \iint_{\|(u,v)-(x,y)\|,r} \frac{u(u)}{f_X(u)} f(u,v)dudv \\
&= \int_{\|v-y\|<r} f_U(v)dy - \iint_{\|(u,v)-(x,y)\|,r} f_U(u,v)dudv \\
&> f_U(y)c_{d_y} r^{d_y}(1 - Cr^2) - f_U(x,y)c_{d_x+d_y} r^{d_x+d_y}(1 + Cr^2) \\
&> f_U(y)c_{d_y} r^{d_y}(1 - \varepsilon/8). \qquad \text{(A.107)}
\end{aligned}
$$

A.2 The Discrete Alphabet Case

Under Assumption 1, we prove a more general version of the theorem. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples drawn from some unknown prior $p_X(x)$ and let $q_X(x)$ be some known distribution on $\mathcal{X}$ such that $q_X(x)/p_X(x) \in [C_3, C_4]$ for all $x \in \mathcal{X}$. Then define

$$
\begin{aligned}
w_x^{(q)} &\equiv \frac{N q_X(x)}{n_x} \ , \qquad \text{(A.108)} \\
n_{y,i}^{(q)} &\equiv \sum_{j \neq i} w_{X_j}^{(q)} \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \qquad \text{(A.109)}
\end{aligned}
$$

The proposed estimator is:

$$
\widehat{I}_{k,N}^{(q)}(X,Y) \equiv \frac{1}{N} \sum_{i=1}^{N} w_{X_i}^{(q)} \Big(\psi(k) + \log(N) - \big(\log(n_{X_i}) + \log(n_{y,i}^{(q)}) \big) \Big) \ . \qquad \text{(A.110)}
$$

We claim that $\widehat{I}^{(q)}_{k,N}$ converges to the true value in probability, i.e.

$$\lim_{N\to\infty} \mathbb{P}\big(\big| \widehat{I}^{(q)}_{k,N}(X,Y) - I^{(q)}(f_{Y|X}) \big| > \varepsilon \big) = 0 \,, \tag{A.111}$$

where

$$I^{(q)}(f_{Y|X}) \;\equiv\; \sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{f_q(y)} dy \tag{A.112}$$

and

$$f_q(y) \;\equiv\; \sum_{x'\in\mathcal{X}} q_x(X) f_{Y|X}(y|x'). \tag{A.113}$$

Notice that Theorem 1 is a special case when $q_X(x)$ is uniform. Define

$$g(X_i, Y_i) \;\equiv\; \psi(k) + \log(N) - \big(\log(n_{x,i}) + \log(n^{(q)}_{y,i}) \big) \,, \tag{A.114}$$

such that $\hat{I}^q_{k,N}(X,Y) = \frac{1}{N}\sum_{i=1}^N w^{(q)}_{X_i} g(X_i, Y_i)$. Define each quantity with the true prior $p_X(x)$ as

$$w'_x \;\equiv\; \frac{q_X(x)}{p_X(x)} \,, \tag{A.115}$$

$$n'_{y,i} \;\equiv\; \sum_{j\neq i} w'_{X_j} \mathbb{I}\{ \|Y_j - Y_i\| < \rho_{k,i} \} \,, \tag{A.116}$$

$$g'(X_i, Y_i) \;\equiv\; \psi(k) + \log(N) - \big(\log(n_{X_i}) + \log(n'_{y,i}) \big) \,. \tag{A.117}$$

We apply triangular inequality, and show that each term converges to zero in probability.

$$
\begin{aligned}
& \big|\widehat{I}^{(q)}_{k,N}(X,Y) - I^{(q)}(f_{Y|X})\big| \\
= \ & \Big|\sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} q_X(x) f_{Y|X}(y|x')} dy \\
& - \frac{1}{N}\sum_{i=1}^{N} w^{(q)}_{X_i} g(X_i,Y_i)\Big| \\
\le \ & \frac{1}{N}\Big| \sum_{i=1}^{N} \big(w^{(q)}_{X_i} g(X_i,Y_i) - w'_{X_i} g'(X_i,Y_i) \big) \Big| \\
& + \Big| \sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} q_X(x) f_{Y|X}(y|x')} dy \\
& - \frac{1}{N}\sum_{i=1}^{N} w'_{X_i} g'(X_i,Y_i) \Big| \, . \qquad\qquad (A.118)
\end{aligned}
$$

The first term (A.118) captures the error in estimating $p_X(x)$. Similar as in (A.22), the probability that it deviates from 0 is is upper bounded by:

$$
\begin{aligned}
& \mathbb{P}\Big(\frac{1}{N}\Big| \sum_{i=1}^{N} \big(w^{(q)}_{X_i} g(X_i,Y_i) - w'_{X_i} g'(X_i,Y_i) \big) \Big| > \varepsilon \Big) \\
\le \ & \mathbb{P}\Big(\max_{x\in\mathcal{X}} |w^{(q)}_x - w'_x| > \varepsilon/(3\log N) \Big) \, , \qquad\qquad (A.119)
\end{aligned}
$$

for sufficiently large N. Recall that $w_x = N q_X(x)/n_x$, (A.119) is bounded by:

$$
\begin{aligned}
& \mathbb{P}\Big(\max_{x\in\mathcal{X}} |w^{(q)}_x - w'_x| > \varepsilon/(3\log N) \Big) \\
= \ & \mathbb{P}\Big(\max_{x\in\mathcal{X}} \Big| \frac{N q_X(x)}{n_x} - \frac{q_X(x)}{p_X(x)} \Big| > \varepsilon/(3\log N) \Big) \\
= \ & \mathbb{P}\Big(\forall x\in\mathcal{X}, n_x \notin [\frac{N q_X(x)}{\frac{q_X(x)}{p_X(x)} + \frac{\varepsilon}{3\log N}}, \frac{N q_X(x)}{\frac{q_X(x)}{p_X(x)} \frac{\varepsilon}{3\log N}}] \Big) \\
\le \ & \mathbb{P}\Big(\forall x\in\mathcal{X}, \Big| n_x - N p_X(x) \Big| > \frac{N\varepsilon p_X^2(x)}{6\log N q_X(x)}) \Big) \, , \qquad (A.120)
\end{aligned}
$$

for sufficiently large N such that $\frac{\varepsilon p_x}{3\log N q_X(x)} < 1/3$. Recall that for each $x\in\mathcal{X}$, $n_x = \sum_{i=1}^{N} \mathbb{I}\{X_i = x\}$. Therefore, n_x is a binomial random variable with parameter $(N, p_X(x))$. Therefore, by Hoeffding's inequality, for any

$x \in \mathcal{X}$, we have:

$$
\begin{aligned}
& \mathbb{P}\big(|n_x - Np_X(x)| > \frac{N\varepsilon p_X^2(x)}{6 \log N q_X(x)} \big) \\
\leq \;\; & 2 \exp\{ -\frac{1}{2N} (\frac{N\varepsilon p_X^2(x)}{6 \log N q_X(x)})^2 \} \\
\leq \;\; & 2 \exp\{ -\frac{N\varepsilon^2 C_1^2}{72|\mathcal{X}|^2 C_4^2 (\log N)^2} \} \,, \qquad \text{(A.121)}
\end{aligned}
$$

where the last inequality comes from the assumption that $p_X(x) > C_1/|\mathcal{X}|$ and $q_X(x)/p_X(x) < C_4$. Then by union bound, (A.120) is upper bounded by:

$$
\begin{aligned}
& \mathbb{P}\Big(\forall x \in \mathcal{X}, \Big| n_x - Np_X(x) \Big| > \frac{N\varepsilon p_X^2(x)}{6 \log N q_X(x)}) \Big) \\
\leq \;\; & |\mathcal{X}| \max_{x \in \mathcal{X}} \mathbb{P}\Big(|n_x - Np_X(x)| > \frac{N\varepsilon p_X^2(x)}{6 \log N q_X(x)} \Big) \\
\leq \;\; & 2|\mathcal{X}| \exp\{ -\frac{N\varepsilon^2 C_1^2}{72|\mathcal{X}|^2 C_4^2 \log N} \}. \qquad \text{(A.122)}
\end{aligned}
$$

Combining with (A.119), we know that (A.118) converges to 0 in probability.

The second term in the error (A.118) comes from the sample noise in density estimation. we decompose our estimator into three terms:

$$
\begin{aligned}
& \frac{1}{N} \sum_{i=1}^{N} w'_{X_i} g'(X_i, Y_i) \\
= \;\; & \widehat{H}_{k,N}^q(Y) - \widehat{H}_{k,N}^q(Y|X) - \sum_{i=1}^{N} \frac{w'_{X_i}}{N} (\log \frac{N-1}{N} + \log(n_{X_i}) - \psi(n_{X_i})) \,, \qquad \text{(A.123)}
\end{aligned}
$$

where

$$
\widehat{H}_{k,N}^q(Y|X) \;\; \equiv \;\; \sum_{i=1}^{N} \frac{w'_{X_i}}{N} \big(-\psi(k) + \psi(n_{X_i}) + \log c_{d_y} + d_y \log \rho_{k,i} \big) \,, \qquad \text{(A.124)}
$$

$$
\widehat{H}_{k,N}^q(Y) \;\; \equiv \;\; \sum_{i=1}^{N} \frac{w'_{X_i}}{N} \big(-\log n_{y,i} + \log(N-1) + \log c_{d_y} + d_y \log \rho_{k,i} \big). \qquad \text{(A.125)}
$$

Notice that $\sum_{i=1}^{N} \frac{w'_{X_i}}{N}(\log(N-1) - \log N + \log(n_{X_i}) - \psi(n_{X_i}))$ converges to 0 in probability as N goes to infinity. The desired claim follows directly from the following two lemmas showing the convergence each entropy estimates to corresponding conditional entropy $H^q(Y|X)$ and entropy $H^q(Y)$. The desired claim immediately follows the two lemmas.

Lemma 8. *Under the hypotheses of Theorem 1, for all $\varepsilon > 0$*

$$\lim_{N\to\infty} \mathbb{P}\left(\left| \widehat{H}^q_{k,N}(Y|X) + \sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_{Y|X}(y|x) dy \right| > \varepsilon \right) = 0 \; . \tag{A.126}$$

Lemma 9. *Under the hypotheses of Theorem 1, for all $\varepsilon > 0$*

$$\lim_{N\to\infty} \mathbb{P}\left(\left| \widehat{H}^q_{k,N}(Y) + \sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_q(y)) dy \right| > \varepsilon \right) = 0 \; , \tag{A.127}$$

where $f_q(y) = \sum_{x\in\mathcal{X}} q_X(x) f_{Y|X}(y|x)$.

A.2.1 Proof of Lemma 8

Define

$$\hat{f}_{Y|X}(Y_i|X_i) = \frac{\exp\{\psi(k) - \psi(n_{X_i})\}}{c_{d_y} \rho_{k,i}^{d_y}} \; , \tag{A.128}$$

so that

$$\hat{H}^q_{k,N}(Y|X) \;\; = \;\; -\sum_{i=1}^{N} \frac{w'_{X_i}}{N} \log \hat{f}_{Y|X}(X_i, Y_i) \; . \tag{A.129}$$

Notice that $\hat{f}_{Y|X}(Y_i|X_i)$ is just the k-nearest-neighbor density estimator for the conditional pdf $f_{Y|X}(y|x)$. Therefore, by Theorem 8 [42], we have

$$\lim_{N\to\infty} \mathbb{E}\big[\log \hat{f}_{Y|X}(Y_i|X_i) \big| (X_i, Y_i) = (x, y) \big] \;\; = \;\; \log f_{Y|X}(y|x) \; . \tag{A.130}$$

Notice that $w'_{X_i} \log \hat{f}(Y_i|X_i)$ are identically distributed, therefore, we have

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}\hat{H}^U_{k,N}(X,Y) \\
= &- \lim_{N\to\infty} \mathbb{E}[w'_{X_i} \log \hat{f}_{X,Y}(X_i,Y_i)] \\
= &- \lim_{N\to\infty} \sum_{x\in\mathcal{X}} \frac{q_X(x)}{p_X(x)} p_X(x) \\
&\Big(\int \mathbb{E}\big[\log \hat{f}_{X,Y}(X_i,Y_i) \big| (X_i,Y_i) = (x,y) \big] f_{Y|X}(y|x) dy \Big) \\
= &- \lim_{N\to\infty} \sum_{x\in\mathcal{X}} q_X(x) \\
&\Big(\int \mathbb{E}\big[\log \hat{f}_{X,Y}(X_i,Y_i) \big| (X_i,Y_i) = (x,y) \big] f_{Y|X}(y|x) dy \Big). \quad \text{(A.131)}
\end{aligned}$$

Use the same technique in the proof of Lemma 2 and Equation (A.47), we can switch the order of limit and integration. Therefore,

$$\begin{aligned}
&\lim_{N\to\infty} \sum_{x\in\mathcal{X}} q_X(x) \Big(\int \mathbb{E}\big[\log \hat{f}_{X,Y}(X_i,Y_i) \big| (X_i,Y_i) = (x,y) \big] f_{Y|X}(y|x) dy \Big) \\
= &\sum_{x\in\mathcal{X}} q_X(x) \Big(\int \lim_{N\to\infty} \mathbb{E}\big[\log \hat{f}_{X,Y}(X_i,Y_i) \big| (X_i,Y_i) = (x,y) \big] f_{Y|X}(y|x) dy \Big) \\
= &\sum_{x\in\mathcal{X}} q_X(x) \int \log_{Y|X}(y|x) f_{Y|X}(y|x) dy \ . \quad \text{(A.132)}
\end{aligned}$$

Therefore,

$$\lim_{N\to\infty} \mathbb{E}\hat{H}^q_{k,N}(Y|X) = -\sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_{Y|X}(y|x) dy. \quad \text{(A.133)}$$

Moreover, by Theorem 11 [42], we have:

$$\lim_{N\to\infty} \mathrm{Var}[\hat{f}_{Y|X}(Y_i|X_i)] \ = \ (\frac{\Gamma'(k)}{\Gamma(k)})' \, \mathrm{Var}[\log f_{Y|X}(y|x)] < \infty \ , \quad \text{(A.134)}$$

and for any $j \neq i$:

$$\lim_{N\to\infty} \mathrm{Cov}[\hat{f}_{Y|X}(Y_i|X_i), \hat{f}_{Y|X}(Y_j|Y_i)] = 0. \quad \text{(A.135)}$$

Since $w'_x \leq C_4$ for all x, similarly as in Lemma 2, we obtain

$$\lim_{N \to \infty} \mathrm{Var}\big[\hat{H}^q_{k,N}(Y|X)\big] = 0\,. \tag{A.136}$$

Combining (A.133) and (A.136), we know $\hat{H}^q_{k,N}(Y|X)$ converges to its mean in L^2, hence in probability, i.e.,

$$\lim_{N \to \infty} \mathbb{P}\Big(\big|\widehat{H}^q_{k,N}(X,Y) + \sum_{x \in \mathcal{X}} q_X(x) \int \log_{Y|X}(y|x) f_{Y|X}(y|x) dy\big| > \varepsilon \Big) = 0\ . \tag{A.137}$$

A.2.2 Proof of Lemma 9

Define

$$\hat{f}_q(Y_i) \;\; \equiv \;\; \frac{n_{y,i}}{(N-1) c_{d_y} \rho_{k,i}^{d_y}}\ , \tag{A.138}$$

such that

$$\hat{H}^q_{k,N}(Y) = -\sum_{i=1}^{N} \frac{w'_{X_i}}{N} \log \hat{f}_q(Y_i). \tag{A.139}$$

By triangle inequality, we can write the formula in Lemma 9 as:

$$\begin{aligned}
& \big|\widehat{H}^q_{k,N}(Y) - \big(-\sum_{x \in \mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_q(y)\big) dy\,\big)\big| \\
= \;\; & |\sum_{i=1}^{N} \frac{w'_{X_i}}{N} \log \hat{f}_q(Y_i) - \sum_{x \in \mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_q(y)\big) dy| \\
\leq \;\; & |\sum_{i=1}^{N} \frac{w'_{X_i}}{N} \log f_q(Y_i) - \sum_{x \in \mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_q(y)\big) dy| \qquad \text{(A.140)} \\
& + \sum_{i=1}^{N} \frac{w'_i}{N} \Big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \Big|. \qquad \text{(A.141)}
\end{aligned}$$

The first term (A.140) is from sampling. Recall that $w'_{X_i} = q_X(X_i)/p_X(X_i)$. Therefore by the strong law of large numbers,

$$\sum_{i=1}^{N} \frac{w'_{X_i}}{N} \log f_q(Y_i) \to \mathbb{E}\Big(\frac{q_X(x)}{p_X(x)} \log f_q(y) \Big) \,, \tag{A.142}$$

almost surely. The mean is given by

$$\mathbb{E}\Big(\frac{q_X(x)}{p_X(x)} \log f_q(y) \Big) \;=\; \sum_{x \in \mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log f_q(y) dy. \tag{A.143}$$

Therefore, (A.140) converges to 0 almost surely.

The second term (A.141) comes from density estimation. For any fixed $\varepsilon > 0$, by union bound, we obtain that

$$\begin{aligned} & \mathbb{P}\Big(\sum_{i=1}^{N} \frac{w'_{X_i}}{N} \big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon \Big) \\ \leq \;& \mathbb{P}\Big(\bigcup_{i=1}^{N} \big\{ \big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon/2 \big\} \Big) + \mathbb{P}\big(\sum_{i=1}^{N} w'_{X_i} > 2N \big) \,. \end{aligned} \tag{A.144}$$

The second term converges to zero by the law of large numbers. The first term is bounded by:

$$\mathbb{P}\Big(\bigcup_{i=1}^{N} \Big\{ \Big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \Big| > \varepsilon/2 \Big\} \Big) \tag{A.145}$$

$$\begin{aligned} \leq \;& N \cdot \mathbb{P}\Big(\Big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \Big| > \varepsilon/2 \Big) \\ \leq \;& N \sum_{x \in \mathcal{X}} p_X(x) \int \big(I_1(x,y) + I_2(x,y) + I_3(x,y) \big) f_{Y|X}(y|x) dy \,, \end{aligned} \tag{A.146}$$

where

$$
\begin{aligned}
I_1(x,y) &= \mathbb{P}\big(\,\rho_{k,i} > r_1 \big| X_i = x, Y_i = y\,\big)\ , && \text{(A.147)}\\
I_2(x,y) &= \mathbb{P}\big(\,\rho_{k,i} < r_2 \big| X_i = x, Y_i = y\,\big)\ , && \text{(A.148)}\\
I_3(x,y) &= \int_{r=r_2}^{r_1} \mathbb{P}\Big(\,\big|\log \hat{f}_q(Y_i) - \log f_q(Y_i)\big| > \varepsilon/2 \\
&\qquad \Big|\rho_{k,i} = r, (X_i,Y_i) = (x,y)\,\Big) f_{\rho_{k,i}}(r) dr\ , && \text{(A.149)}
\end{aligned}
$$

where $f_{\rho_{k,i}}(r)$ is the pdf of $\rho_{k,i}$ given X_i and Y_i. Here

$$
\begin{aligned}
r_1 &= (N^{1/2} p_X(x) f_{Y|X}(y|x) c_{d_y})^{-\frac{1}{d_y}} && \text{(A.150)}\\
r_2 &= (\log N)^{1+\delta/2} (N f_q(y) c_{d_y})^{-\frac{1}{d_y}}. && \text{(A.151)}
\end{aligned}
$$

We will consider the three terms separately.

I_1: Let $B(x,y,r) = \{(X,Y) : \|Y-y\| < r, X = x\}$ be the d_y-dimensional ball centered at y with radius r with same x. Since the Hessian matrix of $H(f_{Y|X})$ exists and $\|H(f_{Y|X})\|_2 < C$ almost everywhere for any $x \in \mathcal{X}$, then for sufficiently small r, the probability mass within $B(x,y,r)$ is given by

$$
\begin{aligned}
&\mathbb{P}\big(\,(u,v) \in B(x,y,r)\,\big) = p_X(x) \int_{\|v-y\|\le r} f_{Y|X}(v) dv \\
=\ & p_X(x) \int_{\|v-y\|\le r} \big(\, f_{Y|X}(y) + (v-y)^T \nabla f_{Y|X}(y) \\
& + (v-y)^T H(f_{Y|X})(y)(v-y)\,\big) dv \\
\in\ & \big[\, p_X(x) f_{Y|X}(y|x) c_{d_y} r^{d_y} (1 - Cr^2)), p_X(x) f_{Y|X}(y|x) c_{d_y} r^{d_y} (1 + Cr^2))\,\big]\ . && \text{(A.152)}
\end{aligned}
$$

Then for sufficiently large N, the probability mass within $B(x,y,r_1)$ is lower bounded by

$$
\begin{aligned}
p_1 &\equiv \mathbb{P}\big(\,(u,v) \in B(x,y,r_1)\,\big) \\
&\ge p_X(x) f_{Y|X}(y|x) c_{d_y} r_1^{d_y} \big(\,1 - Cr_1^2\,\big) \\
&\ge \frac{1}{2} N^{-1/2}. && \text{(A.153)}
\end{aligned}
$$

$I_1(x,y)$ is the probability that at most k samples fall in $B(x,y,r_1)$, so it is upper bounded by

$$
\begin{aligned}
I_1(x,y) &= \mathbb{P}\big(\,\rho_{k,i} > r_1 \big| (X_i, Y_i) = (x,y)\,\big) \\
&= \sum_{m=0}^{k-1} \binom{N-1}{m} p_1^m (1-p_1)^{N-1-m} \\
&\leq \sum_{m=0}^{k-1} N^m (1-p_1)^{N-1-m} \\
&\leq kN^{k-1}(1-\frac{1}{2\sqrt{N}})^{N-k-1} \\
&\leq kN^{k-1}\exp\{-\frac{N-k-1}{2\sqrt{N}}\}\ , \qquad \text{(A.154)}
\end{aligned}
$$

for any $d_x, d_y \geq 1$.

I_2: Let $r_2 = (\log N)^{1+\delta/2}(Nf_q(y)c_{d_y})^{-1/d_y}$. Then for sufficiently large N, the probability mass within $B(x,y,r_2)$ is given by:

$$
\begin{aligned}
p_2 &\equiv \mathbb{P}\big(\,u \in B(x,y,r_2)\,\big) \\
&\leq p_X(x) f_{Y|X}(y|x) c_{d_y} r_2^{d_y}\big(\,1 + Cr_2^2\,\big) \\
&\leq \frac{2p_X(x)f_{Y|X}(y|x)}{f_q(y)} (\log N)^{(1+\delta/2)d_y} N^{-1} \\
&\leq \frac{2p_X(x)f_{Y|X}(y|x)}{\sum_{x\in\mathcal{X}} q_X(x) f_{Y|X}(y|x)} (\log N)^{(1+\delta/2)d_y} N^{-1} \\
&\leq \frac{2}{C_3|\mathcal{X}|N} (\log N)^{(1+\delta/2)d_y}\ , \qquad \text{(A.155)}
\end{aligned}
$$

where the last equation comes from the assumption that $q_X(x)/p_X(x) > C_3$. $I_2(x,y)$ is the probability that at least k samples lying in $B(x,y,r_2)$.

Therefore, it is upper bounded by

$$
\begin{aligned}
I_2(x,y) &= \mathbb{P}\big(\rho_{k,i} < r_2 \big| (X_i,Y_i) = (x,y)\big) \\
&= \sum_{m=k}^{N-1} \binom{N-1}{m} p_2^m (1-p_2)^{N-1-m} \\
&\le \sum_{m=k}^{N-1} \frac{N^m p_2^m}{m!} \\
&\le \sum_{m=k}^{N-1} \frac{N^m p_2^m}{(m/e)^m} \\
&\le \sum_{m=k}^{N-1} (\frac{Nep_2}{k})^m \\
&\le \sum_{m=k}^{N-1} (\frac{2e}{C_3|\mathcal{X}|}(\log N)^{(1+\delta/2)d_y}/k)^m. \qquad \text{(A.156)}
\end{aligned}
$$

Here we use the fact that $m! > (m/e)^m$ for all m. Since $k > (\log N)^{(1+\delta)d_y}$ by assumption, $(\log N)^{(1+\delta/2)d_y}/k$ is decreasing as N increases. For sufficiently large N such that $\frac{2e}{C_1|\mathcal{X}|}(\log N)^{(1+\delta/2)d_y}/k < 1/2$, we obtain:

$$
\begin{aligned}
I_2(x,y) &\le 2(\frac{2e}{C_3|\mathcal{X}|}(\log N)^{(1+\delta/2)d_y}/k)^k \\
&\le 2(\frac{2e}{C_3|\mathcal{X}|})^{(\log N)^{(1+\delta)d_y}}(\log N)^{-\delta(logN)^{(1+\delta)d_y}/2}. \qquad \text{(A.157)}
\end{aligned}
$$

I_3: Given that $(X_i,Y_i) = (x,y)$ and $\rho_{k,i} = r$. Recall that $\hat{f}_q(Y_i) =$

$\frac{n_{y,i}}{(N-1)c_{d_y}r^{d_y}}$, then we have

$$\begin{aligned}
& \mathbb{P}\big(\big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon/2 \big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) \\
= & \mathbb{P}\big(\big| \log n_{y,i} - \log(N-1) - \log c_{d_y} - d_y \log \rho_{k,i} - \log f_q(y) \big| > \varepsilon/2 \\
& \big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) \\
= & \mathbb{P}\big(\big| \log n_{y,i} - \log(N-1) c_{d_y} r^{d_y} f_q(y) \big| > \varepsilon/2 \\
& \big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) \\
= & \mathbb{P}\big(n_{y,i} > (N-1) c_{d_y} r^{d_y} f_q(y) e^{\varepsilon/2} \big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) \\
& + \mathbb{P}\big(n_{y,i} < (N-1) c_{d_y} r^{d_y} f_q(y) e^{-\varepsilon/2} \big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) \; .
\end{aligned} \tag{A.158}$$

Following a similar technique as the analysis of I_4 in proof of Lemma 3, we obtain

$$\begin{aligned}
& \mathbb{P}\big(\big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon/2 \big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) \\
\leq \; & 2 \exp\{ - \frac{C_3 \varepsilon^2}{128(1 + 7\varepsilon/24)} (N-k-1) c_{d_y} r^{d_y} f_q(y) \} \; ,
\end{aligned} \tag{A.159}$$

where C_3 is the lower bound of $q_X(x)/p_X(x)$. Therefore, $I_3(x,y)$ is upper bounded by:

$$\begin{aligned}
I_3(x,y) \; = \; & \int_{r=r_2}^{r_1} \mathbb{P}\big(\big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon/2 \\
& \Big| \rho_{k,i} = r, (X_i, Y_i) = (x,y) \big) f_{\rho_{k,i}}(r) dr \\
\leq \; & \int_{r=r_2}^{r_1} 2 \exp\{ - \frac{C_3 \varepsilon^2}{128(1+7\varepsilon/24)} (N-k-1) c_{d_y} r^{d_y} f_q(y) \} f_{\rho_{k,i}}(r) dr \\
\leq \; & 2 \exp\{ - \frac{C_3 \varepsilon^2}{256} N c_{d_y} f_q(y) ((\log N)^{1+\delta/2} (N f_q(y) c_{d_y})^{-\frac{1}{d_y}})^{d_y} \} \\
\leq \; & 2 \exp\{ - \frac{C_3 \varepsilon^2}{256} (\log N)^{(1+\delta/2) d_y} \} \; ,
\end{aligned} \tag{A.160}$$

for sufficiently large N such that $(N-k-1)/(1+7\varepsilon/24)) > N/2$.

Now combine (A.154), (A.157), and (A.160), and we obtain

$$
\begin{aligned}
& \mathbb{P}\Big(\sum_{i=1}^{N} \frac{w_i'}{N} \big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon \Big) \\
\leq \;& N \sum_{x \in \mathcal{X}} p_X(x) \int (I_1(x,y) + I_2(x,y) + I_3(x,y)) dy \\
\leq \;& kN^k \exp\{-\frac{N-k-1}{2\sqrt{N}}\} + 2N \exp\{-\frac{C_3 \varepsilon^2}{256} (\log N)^{(1+\delta/2)d_y}\} \\
+ \;& 2N (\frac{2e}{C_3|\mathcal{X}|})^{(\log N)^{(1+\delta)d_y}} (\log N)^{-\delta(logN)^{(1+\delta)d_y}/2}. \qquad \text{(A.161)}
\end{aligned}
$$

One can easily see that the first and second terms converges to 0 as N goes to infinity, given that $k < \sqrt{N}/(5 \log N)$. To see that the last term converges to 0, we will show that the logarithm goes to $-\infty$ as N goes to infinity, which is

$$
\begin{aligned}
& \log(N(\frac{2e}{C_3|\mathcal{X}|})^{(\log N)^{(1+\delta)d_y}} (\log N)^{-\delta(logN)^{(1+\delta)d_y}/2}) \\
= \;& \log N + \log(\frac{2e}{C_3|\mathcal{X}|})(\log N)^{(1+\delta)d_y} - \log N \delta (\log N)^{(1+\delta)d_y}/2 \\
= \;& \log N + \log(\frac{2e}{C_3|\mathcal{X}|})(\log N)^{(1+\delta)d_y} - \frac{\delta}{2} (\log N)^{(1+\delta)d_y+1}. \qquad \text{(A.162)}
\end{aligned}
$$

The negative term has the larger exponent, so the logarithm will goes to $-\infty$, and we have

$$
\lim_{N \to \infty} \mathbb{P}\Big(\sum_{i=1}^{N} \frac{w_{X_i}'}{N} \big| \log \hat{f}_q(Y_i) - \log f_q(Y_i) \big| > \varepsilon \Big) = 0. \qquad \text{(A.163)}
$$

Therefore, by combining the convergence of error from sampling and error from density estimation, we obtain that $\widehat{I}_{k,N}^{(q)}(X,Y)$ converges to $I^{(q)}(f_{Y|X})$ in probability.

APPENDIX B

PROOF OF THEOREM 2

Assumption 2. *We make the following assumptions:*

(a) $\int f_{Y|X}(y|x)\big|\log f_{Y|X}(y|x)\big|dy < \infty$, *for all* $x \in \mathcal{X}$.

(b) $\int f_{Y|X}(y|x)\big(\log f_{Y|X}(y|x)\big)^2 dy < \infty$, *for all* $x \in \mathcal{X}$.

(c) *There exists a finite constant* C *such that the Hessian matrix of* $H(f_{Y|X})$ *exists and* $\|H(f_{Y|X})\|_2 < C$ *almost everywhere, for all* $x \in \mathcal{X}$.

(d) *There exists a finite constant* C' *such that the conditional pdf* $f_{Y|X}(y\big|x) < C'$ *almost everywhere, for all* $x \in \mathcal{X}$.

(e) *There exists finite constants* $C_1 < C_3 < C_4 < C_2$ *such that the ratio of the optimal prior* q^* *of the maximizer in the definition of* $\mathcal{C}(f_{Y|X})$ *and the true prior satisfies that* $q^*_X(x)/p_X(x) \in [C_3, C_4]$ *for every* $x \in \mathcal{X}$.

(e) *There exists finite constants* $C_5 < C_6$ *such that* $p_X(x) > C_5/|\mathcal{X}|$ *and* $p_X(x) < C_6/|\mathcal{X}|$, *for all* $x \in \mathcal{X}$.

Define

$$I(f_{Y|X})(q_X) \equiv \sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} q_X(x') f_{Y|X}(y|x')} dy \ , \tag{B.1}$$

and

$$\hat{I}_{k,N}(X,Y)(w) \equiv \frac{1}{N}\sum_{i=1}^{N} w_{X_i}\Big(\psi(k) + \log(N) - \big(\log(n_{X_i}) + \log(n_{y,i})\big)\Big) \ , \tag{B.2}$$

such that $C(f_{Y|X}) = \max_{q_X \in Q} I(f_{Y|X})(q_X)$ and

$$\hat{C}^{\Delta}_{k,N}(X,Y) = \max_{w \in T_\Delta} \hat{I}_{k,N}(X,Y)(w). \tag{B.3}$$

First, consider the quantity:

$$C^{\Delta}(f_{Y|X}) \;\; \equiv \;\; \max_{q_X \in T_\Delta(Q)} I(f_{Y|X})(q_X) \, , \tag{B.4}$$

where the constraint set $T_\Delta(Q)$ is defined as:

$$\begin{aligned} & T_\Delta(Q) = \{q_X \in \mathbb{R}^{|\mathcal{X}|} : [(q_X(x)/p_X(x))] \in T_\Delta \, , \\ \text{and} \qquad & \sum_{x \in \mathcal{X}} q_X(x) \in [1 - |\mathcal{X}|\Delta, 1 + |\mathcal{X}|\Delta]\} \, . \end{aligned} \tag{B.5}$$

We rewrite the error term in Theorem 4.2 as

$$\left|\widehat{\mathcal{C}}^{\Delta}_{k,N}(X,Y) - C(f_{Y|X})\right| \;\; \leq \;\; \left|C^{\Delta}(f_{Y|X}) - C(f_{Y|X})\right| + \left|\widehat{\mathcal{C}}^{\Delta}_{k,N} - C^{\Delta}(f_{Y|X})\right| \, . \tag{B.6}$$

The first error comes from quantization. Let q^* be the maximizer of $C(f_{Y|X})$. By the assumptions, $q^*(x)/p_X(x) \in [C_3, C_4] \subseteq [C_1, C_2]$, for all x. Since $T_\Delta(Q)$ is a quantization of the simplex Q, so there exists a $q_0 \in T_\Delta(Q)$ such that $|q_0(x) - q^*(x)| < \Delta \cdot p_X(x) < \Delta$ for all $x \in \mathcal{X}$. Now we will bound the difference of $I(f_{Y|X})(q_0)$ and $I(f_{Y|X})(q^*)$ by the following lemma.

Lemma 10. *Under the assumptions of Theorem 4.2, if $q(x)/p(x) \in [C_1, C_2]$ and $q'(x)/p(x) \in [C_1, C_2]$ for all $x \in \mathcal{X}$, then*

$$\left| \, I(f_{Y|X})(q) - I(f_{Y|X})(q') \, \right| \leq L \max_{x \in \mathcal{X}} |q(x) - q'(x)| \, , \tag{B.7}$$

for some positive constant L.

Then we have:

$$\begin{aligned} C(f_{Y|X}) &= I(f_{Y|X})(q^*) \\ &\leq I(f_{Y|X})(q_0) + L \max_{x \in \mathcal{X}} |q_0(x) - q^*(x)| \\ &\leq \max_{q \in T_\Delta(Q)} I(f_{Y|X})(q) + L\Delta \\ &= C^{\Delta}(f_{Y|X}) + L\Delta. \end{aligned} \tag{B.8}$$

Similarly, let q^{**} be the maximizer of $C^{\Delta}(f_{Y|X})$, we can also find a $q_1 \in Q$ such that $|q_1(x) - q^{**}(x)| < \Delta$ for all $x \in \mathcal{X}$. Using Lemma 10 again, we will obtain $C^{\Delta}(f_{Y|X}) \leq C(f_{Y|X}) + L\Delta$. Therefore, the first term in (B.6) is bounded by $O(\Delta)$.

Now consider the second term. Upper bound on the second term relies on the convergence of discrete UMI estimation from Theorem 1 . Recall that in the proof of Theorem 1, we have shown that under certain conditions,

$$\mathbb{P}\big(\big|\hat{I}_{k,N}(X,Y)(w_q) - I(f_{Y|X})(q)\big| > \varepsilon/2 \big) \stackrel{N\to\infty}{\longrightarrow} 0 \tag{B.9}$$

for any q with bounded q_X/p_X. Here $(w_q)_x = q(x)/p_X(x)$. Since the set $T_\Delta(Q)$ is finite, by union bound, we have:

$$\begin{aligned}
& \lim_{N\to\infty} \mathbb{P}\big(\forall q \in T_\Delta(Q), \big|\hat{I}_{k,N}(X,Y)(w_q) - I(f_{Y|X})(q)\big| \leq \varepsilon/2 \big) \\
\geq\ & 1 - |T_\Delta(Q)| \lim_{N\to\infty} \mathbb{P}\big(\big|\hat{I}^{k,N}(X,Y)(w_q) - I(f_{Y|X})(q)\big| \leq \varepsilon/2 \big) \\
=\ & 1.
\end{aligned} \tag{B.10}$$

Also, by the strong law of large numbers, we have that

$$\lim_{N\to\infty} \mathbb{P}\big(\forall x \in \mathcal{X}, |p_x(X) - n_X/N| < \Delta/C_2|\mathcal{X}| \big) = 1. \tag{B.11}$$

We claim that if the events inside the probability in (B.10) and (B.11) happen simultaneously, then $\big|\widehat{\mathcal{C}}^{\Delta}_{k,N} - C^{\Delta}(f_{Y|X})\big| < \varepsilon + O(\Delta)$, which implies the desired claim.

Let $w^* = \arg\max_{w \in T_\Delta} \hat{I}_{k,N}(X,Y)(w)$. Define $q_2(x) = w^*_x p_X(x)$. Since $[q_2(x)/p_X(x)] \in T_\Delta$ for all x and

$$\begin{aligned}
\big| \sum_{x\in\mathcal{X}} q_2(x) - 1 \big| &= \big| \sum_{x\in\mathcal{X}} w^*_x (p_X(x) - n_x/N) \ + (\Delta/2)|\mathcal{X}|\big| \\
&\leq |\mathcal{X}| \left((\Delta/2) + C_2 \max_{x\in\mathcal{X}} \big| p_X(x) - n_x/N \big|\right) \\
&\leq (|\mathcal{X}|/2 + 1)\Delta\ .
\end{aligned} \tag{B.12}$$

Therefore, $q_2 \in T_\Delta(Q)$, so

$$\begin{aligned}
\widehat{\mathcal{C}}^{\Delta}_{k,N} &= \hat{I}_{k,N}(X,Y)(w^*) \\
&\leq I(f_{Y|X})(q_2) + \varepsilon \\
&\leq C^{\Delta}(f_{Y|X}) + \varepsilon.
\end{aligned} \tag{B.13}$$

On the other hand, consider $q^{**} = \arg\max_{q_X \in T_\Delta(Q)} I(f_{Y|X})(q_X)$ again, and define $(w_0)_x = q^{**}(x)/p_X(x)$. We know that $(w_0) \in T_\Delta^{|\mathcal{X}|}$ but not necessarily $\sum_{i=1}^N (w_0)_{X_i} = N$. But we claim that the sum is closed to N as follows

$$\begin{aligned}
\Big|\sum_{i=1}^N (w_0)_{X_i} - N\Big| &= \Big|\sum_{x\in\mathcal{X}} \frac{n_x q^{**}(x)}{p_X(x)} - N\Big| \\
&\leq N \max_{x\in\mathcal{X}} \Big\{ \frac{q^{**}(x)}{p_X(x)} \Big| \frac{n_x}{N} - p_X(x) \Big| \Big\} \\
&\leq N C_2 \frac{\Delta}{C_2|\mathcal{X}|} < N\Delta\,.
\end{aligned} \tag{B.14}$$

So we can find a $(w_1) \in T_\Delta(W)$ such that $|(w_1)_x - (w_0)_x| \leq \Delta$ for all x. Let $q_4(x) = (w_1)_x p_X(x)$, similar as (B.12), we know that $q_4 \in T_\Delta(Q)$. Moreover, $\big| q_4(x) - q^{**}(x) \big| \leq p_X(x) \big| (w_1)_x - (w_0)_x \big| \leq \Delta$ for all x. Then we have

$$\begin{aligned}
C^{\Delta}(f_{Y|X}) &= I(f_{Y|X})(q^{**}) \\
&\leq I(f_{Y|X})(q_4) + L \max_{x\in\mathcal{X}} |q^{**}(x) - q_4(x)| \\
&\leq \hat{I}_{k,N}(X,Y)(w_1) + \varepsilon + L\Delta \\
&= \hat{C}^{\Delta}_{k,N} + \varepsilon + L\Delta.
\end{aligned} \tag{B.15}$$

Therefore, we have $|\hat{C}^{\Delta}_{k,N} - C(f_{Y|X})| < \varepsilon + O(\Delta)$, thus our proof is complete.

We give the proof of Lemma 10 below. We will show that for any $x \in \mathcal{X}$, we have $|\frac{\partial}{\partial q_X(x)} I(f_{Y|X})(q)| \leq L/|\mathcal{X}|$ for some L. Therefore,

$$\begin{aligned}
\big| I(f_{Y|X})(q) - I(f_{Y|X})(q') \big| &\leq \sum_{x\in\mathcal{X}} \Big|\frac{\partial I(f_{Y|X})(q)}{\partial q_X(x)}\Big| \, |q_X(x) - q'_X(x)| \\
&\leq L \max_{x\in\mathcal{X}} |q_X(x) - q'_X(x)|\,.
\end{aligned} \tag{B.16}$$

Let $f_q(y) = \sum_{x\in\mathcal{X}} q_X(x) f_{Y|X}(y|x)$. Since $q_X(x) \in [C_1 p_X(x), C_2 p_X(x)] \subseteq$

$[C_1C_5/|\mathcal{X}|, C_2C_6/|\mathcal{X}|]$ we know that

$$f_q(y) \in [C_1C_5 \min_{x\in\mathcal{X}} f_{Y|X}(y|x), C_2C_6 \max_{x\in\mathcal{X}} f_{Y|X}(y|x)] \ , \tag{B.17}$$

for all x, y. Therefore, the absolute value of the gradient is bounded by

$$\begin{aligned}
&\Big| \frac{\partial I(f_{Y|X})(q)}{\partial q_X(x)} \Big| \\
= &\ \Big|\frac{\partial}{\partial q_X(x)}\Big(\sum_{x\in\mathcal{X}} q_X(x) \int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{f_q(y)} dy \Big)\Big| \\
= &\ \Big|\int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{f_q(y)} dy\Big| \\
&+ \Big|\sum_{x'\in\mathcal{X}} q_X(x') \int \frac{f_{Y|X}(y|x') f_{Y|X}(y|x)}{f_q(y)} dy\Big| \\
\leq &\ \Big|\max_y \log \frac{f_{Y|X}(y|x)}{f_q(y)}\Big| + \Big|\max_y \frac{f_{Y|X}(y|x)}{f_q(y)}\Big| \\
\leq &\ \max\{|\log C_1C_5|, |\log C_2C_6|\} + 1/(C_1C_5) \ ,
\end{aligned} \tag{B.18}$$

where $L = |\mathcal{X}| \max\{|\log C_1C_5|, |\log C_2C_6|\}$.

APPENDIX C

PROOF OF PROPOSITION 2

The proof steps are similar to that of Proposition 1, only requiring citations to properties of Rényi divergence and asymmetric information.

- Clearly Axiom 0 holds — it follows from a standard result that $D_\lambda = 0$ if and only if $P = Q$ almost everywhere [37].

- Axiom 1: Suppose $\mathrm{CMI}_\lambda(P_{Z|X})$ is achieved with P_X^*. Consider the joint distribution $P_X^* P_{Y|X} P_{Z|Y}$. Utilizing the data-processing inequality for asymmetric mutual information (cf. Equation (55) in [43]), we get

$$\begin{aligned} \mathrm{CMI}_\lambda(P_{Y|X}) &= \max_{P_X} K_\lambda(P_X P_{Y|X}) \geq K_\lambda(P_X^* P_{Y|X}) \\ &\geq K_\lambda(P_X^* P_{Z|X}) = \mathrm{CMI}_\lambda(P_{Z|X}). \end{aligned} \tag{C.1}$$

  Thus Axiom 1a is satisfied. Now consider Axiom 1b. With the same joint distribution, let P_Y^* be the marginal of Y. Then we have

$$\begin{aligned} \mathrm{CMI}_\lambda(P_{Z|Y}) &= \max_{P_Y} K_\lambda(P_Y P_{Z|Y}) \geq K_\lambda(P_Y^* P_{Z|Y}) \\ &\geq K_\lambda(P_X^* P_{Z|X}) = \mathrm{CMI}_\lambda(P_{Z|X}). \end{aligned} \tag{C.2}$$

- Axiom 2: The asymmetric mutual information has the same additivity property as traditional mutual information, cf. Theorem 27 of [44]. The corresponding additivity for CMI_λ now follows.

- Axiom 3a: The information-centroid representation for CMI_λ states that (see [37] or Equation (44) of [43]):

$$\mathrm{CMI}_\lambda(P_{Y|X}) = \min_{Q_Y} \max_{x} D_\lambda(P_{Y|X=x} \| Q_Y). \tag{C.3}$$

  This characterization allows us to make the observation that CMI_λ is a function only of the convex hull of the probability distributions

$P_{Y|X=x}$, just as earlier: given a conditional probability distribution $P_{Y|X}$, we augment the input alphabet to have one more input symbol x' such that $P_{Y|X=x'} = \sum_x \alpha_x P_{Y|X=x}$ is a convex combination of the other conditional distributions. We claim that the CMI_λ of the new channel is unchanged: one direction is obvious, i.e., the new channel has capacity greater than or equal to the original channel, since adding a new symbol cannot decrease capacity. To show the other direction, we use (2.5) and observe that, due to the quasi convexity of Rényi divergence in its arguments (cf. Theorem 13 in [44]), we get,

$$D_\lambda(P_{Y|X=x'} \| q_Y) = D_\lambda(\sum_x \alpha_x P_{Y|X=x} \| q_Y) \leq \max_x D_\lambda(P_{Y|X=x} \| q_Y).$$

Thus CMI_λ is only a function of the convex hull of the range of the map $P_{Y|X}$, satisfying Axiom 3a. This function is monotonic directly from (C.3), thus satisfying Axiom 3b.

- Axiom 4: For fixed output alphabet $\mathcal{Y}$, it is clear that $\max_{\mathcal{X}, P_{Y|X}} \mathrm{CMI}_\lambda = \log |\mathcal{Y}|$ for each λ. Now suppose for some conditional distribution $P_{Y|X}$ we have $\mathrm{CMI}_\lambda(P_{Y|X}) = \log |\mathcal{Y}|$. This implies that, with the optimizing input distribution, $H_\lambda(Y) - H_\lambda(Y|X) = \log |\mathcal{Y}|$. This implies that $H_\lambda(Y) = \log |\mathcal{Y}|$ and $H_\lambda(Y|X) = 0$, thus Y is a deterministic function of the essential support of X and since $H_\lambda(Y) = \log |\mathcal{Y}|$, the Schur concavity of Rényi entropy (cf. Theorem 1 of [45]) implies that $P_Y = U_Y$, the uniform distribution and the deterministic function is onto.

REFERENCES

[1] C. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423 and 623–656, 1948.

[2] A. Rényi, "On measures of dependence," *Acta Mathematica Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[3] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008.

[4] J. Pearl, *Causality*. Cambridge University Press, 2009.

[5] R. J. E. T. S. Richardson and U. Tutorial, "Non-parametric causal models," 2015.

[6] J. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: Methods and benchmarks," *Journal of Machine Learning Research*, 2015.

[7] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, B. Schölkopf et al., "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[8] S. Krishnaswamy, M. Spitzer, M. Mingueneau, S. Bendall, O. Litvin, E. Stone, D. Pe'er, and G. Nolan, "Conditional density-based analysis of T cell signaling in single-cell data," *Science*, vol. 346, no. 6213, p. 1250689, 2014.

[9] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[10] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[11] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–248, 1989.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, 2012.

[13] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial.* Now Publishers Inc, 2004.

[14] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, p. 026209, 2007.

[15] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[16] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via-nearest-neighbor distances," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2392–2405, 2009.

[17] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems*, 2010, pp. 1849–1857.

[18] K. Sricharan, R. Raich, and A. O. Hero III, "Empirical estimation of entropy functionals with confidence," *arXiv preprint arXiv:1012.4188*, 2010.

[19] B. Póczos, L. Xiong, and J. Schneider, "Nonparametric divergence estimation with applications to machine learning on distributions," *arXiv preprint arXiv:1202.3758*, 2012.

[20] S. Gao, G. V. Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," *arXiv preprint arXiv:1411.2003*, 2014.

[21] S. Gao, G. V. Steeg, and A. Galstyan, "Estimating mutual information by local Gaussian approximation," *arXiv preprint arXiv:1508.00536*, 2015.

[22] K. Kandasamy, A. Krishnamurthy, B. Poczos, and L. Wasserman, "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, 2015, pp. 397–405.

[23] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[24] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *arXiv preprint arXiv:1604.03006*, 2016.

[25] J. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert, "Adaptive multiple importance sampling," *Scandinavian Journal of Statistics*, vol. 39, no. 4, pp. 798–812, 2012.

[26] A. B. Tsybakov and E. Van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scandinavian Journal of Statistics*, pp. 75–83, 1996.

[27] L. Devroye and C. S. Penrod, "The consistency of automatic kernel density estimates," *The Annals of Statistics*, pp. 1231–1249, 1984.

[28] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

[29] J. Mueller, T. Jaakkola, and D. Gifford, "Modeling trends in distributions," *arXiv preprint arXiv:1511.04486*, 2015.

[30] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *Information Theory, IEEE Transactions on*, vol. 18, no. 4, pp. 460–473, 1972.

[31] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *Information Theory, IEEE Transactions on*, vol. 18, no. 1, pp. 14–20, 1972.

[32] D. Lombardi and S. Pant, "Nonparametric k-nearest-neighbor entropy estimator," *Physical Review E*, vol. 93, no. 1, p. 013310, 2016.

[33] C. R. Loader et al., "Local likelihood density estimation," *The Annals of Statistics*, vol. 24, no. 4, pp. 1602–1618, 1996.

[34] N. L. Hjort and M. Jones, "Locally parametric nonparametric density estimation," *The Annals of Statistics*, pp. 1619–1647, 1996.

[35] D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf, *Justifying Information-Geometric Causal Inference.* Springer International Publishing, 2015, ch. 18, pp. 253–265.

[36] N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve, "Telling cause from effect in deterministic linear dynamical systems," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR, 2015, pp. 285–294.

[37] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *Information Theory, IEEE Transactions on*, vol. 41, no. 1, pp. 26–34, 1995.

[38] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer, 1990, pp. 303–305.

[39] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[40] P.-O. Amblard and O. J. Michel, "On directed information theory and Granger causality graphs," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 7–16, 2011.

[41] A. Owen, "Monte Carlo theory, methods and examples (book draft)," 2014.

[42] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *American Journal of Mathematical and Management Sciences*, vol. 23, no. 3-4, pp. 301–321, 2003.

[43] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and rényi divergence," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 1327–1333.

[44] T. Van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *Information Theory, IEEE Transactions on*, vol. 60, no. 7, pp. 3797–3820, 2014.

[45] S.-W. Ho and S. Verdú, "Convexity/concavity of Rényi entropy and α-mutual information," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 745–749.